Statistical Methods for Wearable Devices with Applications to Epidemiological Studies

by

Jiawei Bai

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2017

Abstract

Monitoring and assessing the level of physical activity has been an important part of research in many public health and medical studies. However, the conventional self-reports of physical activity were found unreliable due to various reasons. Recent advances of wearable computing technology enabled researchers to deploy accelerometers in health studies as a physical activity assessment tool. Such devices are able to provide objective and continuous measurement of physical activity for as long as a few months. Common types of data collected by accelerometers include high-frequency tri-axial acceleration time series (raw data) and summarized metrics in epochs (count data). Size of the data and uniqueness of the data structure called for development of new statistical methods, which include topics such as analysis of raw or count data and processing of raw data. The purpose of this dissertation is to provide solutions to three types of questions in accelerometry research. First, a dictionary-based classification method was proposed to predict the type of physical activity performed by elderly adults using raw accelerometry data. The classification method decomposed movements into short components called "movelets and built a reference for each

ABSTRACT

activity type. Unknown activities were predicted by matching new movelets to the reference. The movelet method was able to identify a variety of household activities including short activities, such as chair-stands. Second, a set of explicit and open-source metrics for physical activity was introduced to summarize raw accelerometry data into count data. One of the metrics, the Activity Index, was compared with several existing summary metrics and showed to be more sensitive to sedentary and light activities and better associated with energy expenditure. Third, a two-stage regression model was proposed to study the association between minute-by-minute activity count and human demographics. The model allows for both time-varying parameters and time-invariant parameters, which helps capture both the transition dynamics between active/inactive periods (Stage 1) and the activity intensity dynamics during active periods (Stage 2).

Readers:

Ciprian M. Crainiceanu (Thesis Advisor, Dept. of Biostatistics, JHSPH)

Vadim V. Zipunnikov (Thesis Co-advisor, Dept. of Biostatistics, JHSPH)

Jennifer A. Schrack (Thesis Committee Member, Dept. of Epidemiology, JHSPH)

Adam P. Spira (Thesis Committee Chair, Dept. of Mental Health, JHSPH)

Mei-Cheng Wang (Alternate, Dept. of Biostatistics, JHSPH)

Sara Benjamin Neelon (Alternate, Dept. of Health, Behavior and Society, JHSPH)

Acknowledgments

I would like to share my gratitude with all the people who have given me the greatest help and support throughout my 5-year study at the Johns Hopkins Bloomberg School of Public Health.

It is my greatest honor to have worked with my advisor, Dr. Ciprian M. Crainiceanu. He dedicated himself in helping me become a good statistician, a good scientific researcher and most importantly, a good person. As the strongest advocate and supporter of mine, he gave me the absolute trust and freedom, so that I could identify and work on the problems I love. I am very glad we plunged into this interesting area of research on wearable devices 7 years ago. My co-advisor, Dr. Vadim V. Zipunnikov, is always ready to offer me help and advice, no matter how busy he is. He is a great example of how a junior faculty is able to keep up with both his research and students.

I am grateful to all my committee members and collaborators for their help and support throughout the development of my dissertation. I truly appreciate their vision and effort to collect the wonderful wearable device data, which allowed me to be one

of first few statisticians working in this direction. Through those discussions with them, I learned how to identify and approach statistical problems using the scientific context behind the data; I learned how to communicate my results to researchers with different backgrounds. All these things are helping me to become a better and better Biostatistician.

Then, to all my dear friends, who have been kindly around for numerous late night chats, fun drives along the countryside, delicious food hunting everywhere and so on. Five years would not have passed so fast if it was not for them. And a special thanks to Dr. Zhenke Wu, who is always available for help and advice during our 7-year overlap at Johns Hopkins. We have learned from and influenced each others in so many different aspects, from research to life.

I am grateful to all my family, especially my mom and dad, who have consistently been offering me the greatest support. They always urge me to do things I enjoy, which eventually encouraged me to finish my doctoral degree. I am also thankful to Aunt Yan and Uncle Jihong, as they were always there when I needed help and advice for school, research and life.

I am indebted my dear wife, Shuang, who is an amazing source of energy and support. I am proud I found someone whom I can share both my work and life with.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 Background 1

1.2 Organization structure 5

2 Predicting type of activity using raw accelerometry data **7**

2.1 Introduction 7

2.2 Methods 14

2.2.1 Definitions 15

2.2.2 Matching and labeling 18

2.2.3 Movement fingerprints and lazy movelets 20
2.2.4 Summary 22
2.3 Application to LIFEmeter data 23
2.3.1 Constructing the dictionary 24
2.3.2 Initial results 25
2.3.3 Refined results 30
2.4 Discussion 32

3 Accelerometry metrics extracted from raw accelerometry data 36

3.1 Introduction 36
3.2 Data collection 40
3.2.1 Study population 40
3.2.2 Data description 41
3.3 Accelerometer metrics definition 47
3.3.1 Time active 49
3.3.1.1 Scalar summaries of time active 51
3.3.1.2 Cumulative relative time active 52
3.3.2 Activity intensity 54
3.3.2.1 Scalar summaries of activity intensity 56
3.3.2.2 Cumulative Relative Activity Intensity 57
3.4 Evaluation of metrics 58
3.4.1 Validation of metrics 58

3.4.2 Association with health outcomes 61
3.5 Discussion 63
3.6 Supplementary Materials 64
3.6.1 More on the analysis of association with health outcomes . . . 64
3.6.2 More on the validation of Activity Intensity 68

4 An activity index for raw accelerometry data and its comparison with other activity metrics 72

4.1 Introduction 72
4.2 Materials and Methods 76
4.2.1 Participants 76
4.2.2 Accelerometry 76
4.2.3 Data collection 77
4.2.4 The new Activity Index 79
4.2.5 Statistical analysis 82
4.2.5.1 Data processing 82
4.2.5.2 Directly comparing AI, AC and ENMO 82
4.2.5.3 Comparing MET prediction performance of AI, AC and ENMO 83
4.3 Results 84
4.3.1 Summary statistics 84
4.3.2 Directly comparing AI, AC and ENMO 86

4.3.3 Comparing MET prediction performance of AI, AC and ENMO 90
4.4 Discussion . . . 92
4.5 Supplementary Materials . . . 96
4.5.1 Definitions and mathematical formula of AI . . . 96
4.5.2 Properties of AI . . . 99
4.5.2.1 Easy implementation . . . 99
4.5.2.2 Additivity . . . 99
4.5.2.3 Rotational Invariance . . . 99

5 A two-stage model for wearable device data 102

5.1 Background . . . 102
5.2 Two-stage Model . . . 105
5.2.1 Notation . . . 105
5.2.2 Model framework . . . 106
5.2.3 Model specification . . . 108
5.3 Estimation . . . 110
5.4 Simulation Study . . . 114
5.5 Application . . . 119
5.6 Discussion . . . 124
5.7 Supplementary Material . . . 126
5.7.1 Large-sample property of the estimators . . . 126

Bibliography **130**

Vita **148**

List of Tables

2.1 A subject-specific dictionary with with A chapters 17
2.2 A list of activities of interest 24
2.3 Comparison of prediction accuracy based on observer-annotated labels and the predicted labels 29
2.4 Table of prediction agreement for both subjects and both visits, using the combined observer labels 31

3.1 Estimated regression coefficients for 8 different models 67

4.1 Summary statistics of 4 different metrics 85

5.1 Mean integrated squared error across t for the simulation 118

List of Figures

2.1 Two segments of accelerometer data 10
2.2 A display of matching an unlabeled movelet $M_i(t^*)$ to 4 chapters in the dictionary 19
2.3 Illustration of the chapter "Standing from Lying" 21
2.4 Observer-defined annotations and predictions for two segments of accelerometer data with several activity types. Curves giving the smallest distance between movelets and each chapter are displayed. 26
2.5 Comparison combined observer annotations and predicted labels . . . 32

3.1 The raw 3-axis accelerometry data for Subject 3208 and Subject 3056 with Active v.s. Inactive prediction results 42
3.2 The density curves of standard deviations for all (34) subjects 46
3.3 Two periods of raw data (panel 1 and 2), TA bars for Subject 3092 (panel 3). 50
3.4 Cumulative Relative Time Active and Cumulative Relative Activity Intensity plots for 3 subjects 53
3.5 Metrics validation results for Subject 3056 and Subject 3092 during two different visits 60
3.6 Scatterplot of 7 covariates versus 4 outcomes in the association study 65
3.7 AI for subject 3106 for normal and brisk walking 68
3.8 AI for subject 3208 for normal and brisk walking 69
3.9 Median and standard deviation of AI for 10 subjects during normal and brisk walking 70

4.1 A general framework for accelerometer-related studies 74
4.2 Scatterplots of Activity Index versus activity count (A) and Activity Index versus Euclidean Norm Minus One (B) for a randomly selected participant 87

4.3 Comparison of the boxplots of Activity Index, activity count, activity count with Low Frequency Extention and Euclidean Norm Minus One during different types of activities 88
4.4 The "receiver operating characteristic" (ROC) curves for distinguishing four pairs of activity types using 4 different metrics 89
4.5 Scatterplots of metabolic equivalents versus 4 different metrics 91
4.6 The "receiver operating characteristic" curves of 4 different metrics to classify between different activity intensity categories based on metabolic equivalents 92

5.1 Illustration of conceptual model structure 106
5.2 Comparison of true coefficient curves, mean estimated coefficient curves and the estimated coefficient curves 115
5.3 Average bias for all time-varying and structural parameters 116
5.4 Estimated covariate coefficients of the two-stage model 121
5.5 Comparison of estimated coefficients of gender as time-varying and time-invariant coefficients 124

Chapter 1

Introduction

1.1 Background

Physical activity has been recognized to be beneficial to human health in various population for a long time [45, 61]. As part of the effort to promote conducting physical activity, an official guideline for physical activity was published in 1995 by Centers for Disease Control and Prevention and American College of Sports Medicine, recommending US adults to conduct at least 150 minutes of moderate to vigorous physical activities (MVPA) per week [65]. Accurate assessment of physical activity is therefore always a crucial part of the research, because one must understand the current levels and changes of physical activity, before assessing the effectiveness of various interventions to increase activity levels.

Conventional methods for measuring physical activity are usually based on self-

reported questionnaires or diaries. These methods are usually easy to implement in large population because of its low participant burden and low cost [27], but are found to be unreliable due to various reasons (issues of recall, response bias, etc.). The increasing need of more valid and unbiased measurement of physical activity has urged the researchers to explore a number of objective measurements, including calorimetry (i.e. doubly labeled water, indirect calorimetry), physiologic markers (i.e. cardiorespiratory fitness, biomarkers) and motion sensors (i.e. accelerometers, heart rate monitors). Many studies have reported the discrepancy between self-reported and objective measures, although there was no consensus on the direction of such difference [71]. Motion sensors have become an important tool for objective measurement of physical activity [94], because they are able to provide a complete profile of physical activity (e.g. minute-by-minute activity level for 24 hours) – more preferable than the self-reports which only offer a snapshot of physical activities (e.g. daily active time, daily sedentary time).

Although the first use of accelerometers as an activity monitor dated back to 1960s, these sensors were not extensively studied until 1980s [48]. Most early models of accelerometers offered a one-dimensional time series in certain epoch, often referred to as "accelerometer counts" or "activity counts" [91]. These counts are summary metrics that quantify the movement of the device during each epoch (e.g. second, minute), and are used as proxies of human physical activity in the same period (in the rest of this dissertation, the phrase "count data" and "summary metrics" will be used

interchangeably). One of the main research questions is to establish an connection between count data and other meaningful measures such as energy expenditure – studies with such purpose are usually called "calibration studies"(see review [91]). Another type of research is concerned of predicting type of physical activity using the count data, which underwent a lot of methodological development by computer scientists, electronic engineers and statisticians (see review [70]). In the past few years, some researchers began to study the relationship between count data and human health without translating counts to energy expenditure or activity types, hoping to mitigate the effect of biases introduced during such translation. With the help of more complicated statistical method, these studies were able to directly estimate the effect of health factors on circadian rhythm of physical activity [37, 76, 93].

Since late 2000, technology has allowed sampling and storing high-resolution raw accelerometry data on newer devices, including ActiGraph GT3X+ (ActiGraph, Pensacola, Florida, USA) [48] and GENEActiv (Activinsights Ltd, Cambridgeshire, UK) [30]. The raw accelerometry data consist of 3 time series corresponding to the sensor's instantaneous acceleration along 3 orthogonal directions, commonly with 10–100 samples per second on each axis. The raw accelerometry data provide high resolution information on the changes of acceleration, and are preferred outputs in modern accelerometry studies [48]. Based on the data, physical activity type recognition algorithms with better performance were proposed [9,13,28,42,73,80]. Statistical and machine learning approaches have also been applied to predict energy expendi-

ture or metabolic equivalents using raw data [28, 80]. The large sample rate even allowed for comprehensive gait analysis [88, 95].

Raw accelerometry data provide rich information on many aspects of human physical activity, but in large public health studies such as US National Health and Nutrition Examination Survey (NHANES) 2011–2014, the size of data sampled at 80 Hz could easily reach over 7 terabytes. Many challenges arise, including the storage, processing and analysis of the data [86]. Moreover, if the population-level circadian rhythm of physical activity level is the major research interest, tri-axial acceleration in milliseconds may not provide much more information than minute-by-minute activity counts. As a result, analytic methods based on the reduced count data are still greatly needed. In the other hand, as the definitions of many versions widely used activity count are proprietary and device- and software-specific [1, 10], a few open-source and reproducible summary metrics have been proposed as replacement of activity count [4, 10, 90]. Further validation studies highlighted their favorable correlation with energy expenditure [8, 44, 89].

The purpose of this dissertation is to provide solutions to several challenges associated with the processing and analysis of accelerometry data.

1.2 Organization structure

This dissertation consists of 5 chapters, each of which was written to be largely self-contained and complete.

Chapter 1 provides an overview of the accelerometry research. We review how accelerometers began to supplement or replace self-reports as assessment tools for physical activity. We then discuss major research topics based on both count data and raw data, followed by introduction of several novel data reduction methods that summarize raw data into count data. Finally, we provide an outline of this dissertation.

In Chapter 2, a dictionary-based classification method was proposed to predict the type of physical activity performed by elderly adults using raw accelerometry data. The classification method decomposed movements into short components called "movelets", and built a reference for each activity type. Unknown activities were predicted by matching new movelets to the reference. The movelet method were able to identify a variety of household activities including short activities, such as chair-stands. This chapter is based on "J. Bai, J. Goldsmith, B. Caffo, T. A. Glass, and C. M. Crainiceanu, 'Movelets:A dictionary of movement,' *Electronic Journal of Statistics*, vol. 6, pp. 559–578,2012" [9].

Chapter 3 introduced a set of explicit metrics for physical activity based on raw accelerometry data. The metrics were based on two major concepts: (i) Time Active, a measure of the length of time when activity is distinguishable from rest and (ii)

Activity Intensity, a measure of the relative amplitude of activity relative to rest. The metrics were validated using in-lab replication studies and an association study with health factors. This chapter is based on "J. Bai, B. He, H. Shou, V. Zipunnikov, T. A. Glass, and C. M. Crainiceanu, 'Normalization and extraction of interpretable metrics from raw accelerometrydata', *Biostatistics*, vol. 15, no. 1, pp. 102–116, 2014" [10].

Chapter 4 introduced the Activity Index by further refining Activity Intensity, to summarize raw accelerometry data into count data. Extensive comparison was conducted between Activity Index, Activity Count and Euclidean Norm Minus One to assess the validity of Activity Index, which was subsequently showed to be more sensitive to sedentary and light activities and better associated with energy expenditure. This chapter is based on "J. Bai, C. Di, L. Xiao, K. R. Evenson, A. Z. LaCroix, C. M. Crainiceanu, andD. M. Buchner, 'An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics', *PLoS ONE*, vol. 11, no. 8, p. e0160644, Aug 2016" [8].

In Chapter 5, a two-stage regression model was proposed to study the association between minute-by-minute activity count and human demographics.The model allows for both time-varying parameters and time-invariant parameters, which helps capture both the transition dynamics between active/inactive periods (Stage 1) and the activity intensity dynamics during active periods (Stage 2). The methods were motivated by and applied to the Baltimore Longitudinal Study of Aging.

Chapter 2

Predicting type of activity using raw accelerometry data

2.1 Introduction

Accurate measurement of physical activity is necessary for understanding the complex relationship between an individual's health outcomes and his or her behavior profile. Unfortunately, standard measures of activity such as questionnaires and diaries are based on self-reporting and are subject to known shortcomings. Moreover, these measures typically offer snapshots of activity and do not reflect the dynamic nature of movement in the real world. Recently, progress in sensor technologies and wearable computing devices have allowed researchers to collect real-time information on movement through the use of accelerometers. In this paper, we propose a method

for predicting activity types, such as walking, standing and sitting, from a multichannel accelerometer designed with widespread deployment in observational studies in mind.

In early years, human activity function is assessed using measures of activities of daily living that depend on retrospective self-report, despite well-documented and substantial measurement error associated with these instruments [32, 58]. The results of these studies were highly impaired by problems associated with self-reported activity data. Wearable sensors started to be deployed into studies, since they allow for unbiased measurement in older populations with cognitive or physical impairment. Moreover, the accuracy of sensors is not effected by differences in sex, race/ethnicity or language, all well known sources of bias in self-reports. This is particularly important in the study of aging populations, both because issues with recall are more severe and because understanding physical activity accurately is central to the study of elderly populations in public health [65]. The use of them to collect activity information in large-scale observational studies took a major step forward with the addition of the ActiGraph to the National Health and Nutrition Examination Survey (NHANES) in 2003 [85]. Many published work have demonstrated these devices' ability to monitor human activity status [6, 16, 19, 29, 39, 40, 51, 64, 73, 92]. Some of them focused on the quantification of total energy expenditure [92] or "activity counts" [51]. However, these devices (often combined with more sophisticated sensors) offer the potential to assess more complex questions regarding real-world function and more refined mea-

sures of specific activity types. Accelerometer, which is the basic of these wearable sensors, were discussed in many literatures because they are capable of accurately collecting adequate data for physical activity monitoring [19, 43, 73]. Since avoiding laying a burden to the subjects is crucial in large scale observational study, developing methods to predict the physical activity using accelerometry data becomes one of our major interests.

We base activity prediction on the idea that movements can be understood in terms of smaller components, which we dub "movelets". Briefly, given accelerometer time series data, we decompose movements into short overlapping segments; these movelets are the elements which make up motions and activities. Using data with known activity labels, movelets are organized by activity type into "chapters", or collections of movelets with the same activity label. Predictions of unknown activity labels are made by finding the closest match, defined in terms of squared error for all acceleration channels, of an unlabeled movelet to those in chapters. Thus we build our method on the intuition that movements with elements that look similar are likely to have the same labels.

Our data are generated using a single accelerometer positioned on the subject's hip at the apex of the left iliac crest. The accelerometer is built on core chip MMA7260Q by *Freescale*$^{\text{TM}}$, and records acceleration in three mutually orthogonal directions for a wide range of sampling frequencies (time points per second) and sensitivities (acceleration per unit of scale). Data were collected during in-laboratory sessions in which

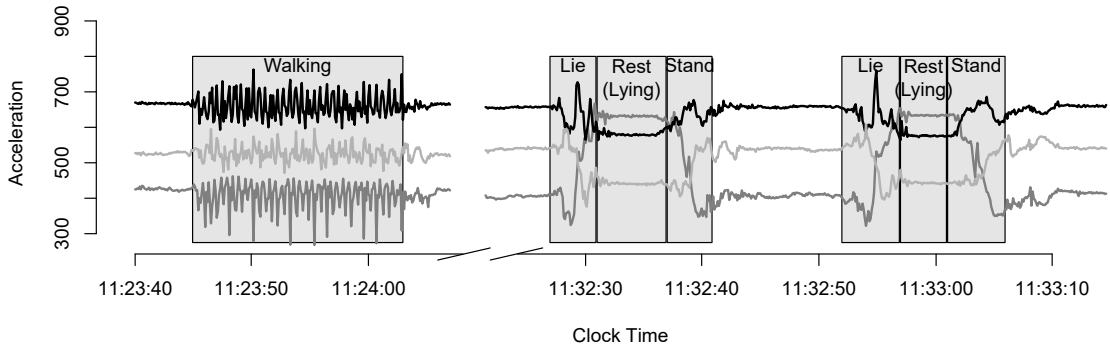


Figure 2.1: Two segments of accelerometer data. First, a subject walks for approximately 20 second; then, a subject preforms two replicates of "Lie down / Rest / Stand up". Acceleration in three mutually orthogonal directions is shown, and activity labels are included.

subjects performed a collection of activities, including resting, walking, and lying, repeated chair stands, lifting an object from the floor, up-and-go, and standing to reclining on a couch. We observe data for two subjects with two laboratory visits each. Sessions lasted roughly 15 to 20 minutes, and in that time each activity was replicated up to three times. Both the data collection device and activities performed are compatible with the needs of observational studies, especially of elderly populations: the single accelerometer worn at the hip is unobtrusive and wearable in real time, and the activities provide a useful understanding of physical movement. During the data collection, an observer recorded activity start and stop times to provide a time series of movement labels that accompanies the accelerometer signal.

The accelerometer output consists of 3 voltage time series, which are proxy measures of acceleration. The time series vary by amplitude, frequency and correlation along the time course of the corresponding activities. For example, Figure 2.1 displays two segments of accelerometer data. In the first segment, the subject stands,

walks twenty meters, and stands. In the second segment, the subject performs two replicates of lying down and standing up; during each replicate, the subject lies from a standing position, rests for several seconds in the lying position, and rises to a standing position. Three acceleration channels or axes are shown, and activity labels are provided. From this figure, we see that active periods, in which the subject is walking, rising or lying down, have higher variability than inactive periods, in which the subject is resting in either the standing or lying position. Walking is characterized by periodic acceleration patterns for each axis, although there are differences in amplitude between axes. Replicates of the "Chair Stand" activity display similar patterns, bolstering the intuition that movements that share a label also appear similar visually. Although there are two types of inactivity (standing and lying), the acceleration time series corresponding to these two periods are characterized by low variation around stable constants; however, the ordering and relative position of the axes are different, due to a change in the orientation of the accelerometer with respect to Earth's gravity.

The goal of this work is to demonstrate the conceptual framework for the movelet approach, rather than to describe the details of its application to a large data set. The movelet prediction algorithm described in the paper is an important first step in developing accelerometer-based biomarkers of activity in large observational studies. Several strengths of this approach are illustrated by the analysis of a few subjects at a few visits; at the same time, improvements and refinements both in the statistical

analysis and in the data collection are suggested by our results. For instance, matching unlabeled movelets to reference chapters provides a fast and easily understood method for predicting labels. However, results can be sensitive to the definition of the gold standard of activity type - very often, the observer annotations disagree with the raw accelerometer output. Additionally, the use of gyroscopic information, which is included in many accelerometer devices, can give accelerometer output that is robust to rotations of the device itself. These are important considerations in designing a data collection method that will give useful information regarding activity in observational studies. More importantly, our findings have already led to changing the proposed design of the experiment for an ongoing and future observational studies. Indeed, an investigator will now go to the home of study participants, help install the device correctly, provide simple hands-on instructions, and ask the participants to perform a few well defined tasks. This process will be videotaped for improving and assisting human annotation. The investigator will then leave and study participants are then called on the phone and asked to perform a few simple tasks for re-calibration. None of these features was part of the original data collection protocol. We conclude that understanding the inherent pitfalls and variability associated with even the most advanced measuring technology can lead to dramatic improvements in the design of experiments, data quality, and analysis. This paper, as a "proof-of-concept" work, provides the first part of the story for accelerometry data.

Prediction of physical activity intensity and type has been under intense method-

ological development in electronic engineering and computer science, but to a lesser extent in statistics. Preece et al. [70] provided a nice review of the current methods of activity prediction. Many prediction methods using either raw or transformed accelerometer data exist, including "cut-point" or linear regression [35, 43], quadratic discriminant analysis [69], artificial neural networks [49, 50, 81, 96, 97], Markov Models [52, 69], unsupervised learning [62] and combined methods [7, 73]. Previous work has often focused on activity types that are not of interest in public health studies [69], such as using computer or brushing of teeth, or has included multiple accelerometers placed at several locations on a subject's body [49, 50, 57]. A comparison of recent approaches was also applied to data generated using five biaxial accelerometers by Bao and Intille [13]. However, these approaches are unsuitable for application to accelerometer data in public health studies, either because they require more sensors on subjects or because they are not designed to detect short-term activities like standing from a lying position. Moreover, prediction results from black-boxed machine learning methods are usually difficult to examine and improve. This stimulates us to find a method which could not only detect long term activities like walking and vacuuming, but also short term activities like sitting down or lying down.

Our approach and taxonomy are inspired by the speech recognition literature [47], where words or parts of words are matched to known speech patterns. However, the parallel with speech recognition should not be overstated given the large differences between the two activities and measurement instruments. First, speech is often

recorded at much higher frequencies (between 8 and 16kHz) than acceleration (10Hz in our dataset), providing density and detail to voice recognition data [67]. Second, audio data is inherently single-channel while acceleration is understood in three orthogonal directions, increasing the dimension of the activity prediction problem. In natural speech most sounds and many full words are repeated often, providing an ample training set on which to build a prediction algorithm. In activity prediction, movements can be rarely performed and infrequently observed, making the definition of a training set challenging. Moreover, high fidelity audio recorders could be treated as thought they were lossless reproductions of the original signal. In contrast, accelerometers are weak proxies for activities that are complex and could be ambiguous.

The remainder of the paper is organized as follows. In Section 2.2 we describe the movelet-based approach to predicting activity based on accelerometer data. Section 2.3 details the application of our proposed method to the real data described above. We close with a discussion in Section 2.4.

2.2 Methods

To predict activities based on accelerometer data, we first define a movelet as a basic element of 3-axis time series data. Collections of movelets paired with known labels (annotations) form chapters, which are in turn organized into reference dictionaries of

known movelets and their associated activities. Classification of accelerometer data with unknown activity annotations is based on decomposing the unlabeled data into component movelets, and then matching each unlabeled movelet to these chapters. The label of the best matched chapter is used as a preliminary prediction of the activity of the unlabeled movelet.

2.2.1 Definitions

We observe data that is a collection of three time series representing the acceleration in three mutually orthogonal axes. Though we have two subjects and each with two visits, we actually treat them as 4 independent visits. Thus denote the data by $\mathbf{X}_i(t) = \{X_{i1}(t), X_{i2}(t), X_{i3}(t)\}$, $t = 1, 2, \ldots, T_i$, where T_i is the length of the accelerometer time series for visit i. Define an activity label time series $L_i(t)$ such that $L_i(t)$ is a function mapping t to $\{\text{Act}_1, \text{Act}_2, \ldots, \text{Act}_A\}$, $t = 1, 2, \ldots, T_i$, where Act_a denotes activity type a. Let $\mathbb{T}_i$ and $\mathbb{V}_i$ be a partition of observation time for visit i into training and validation sets, respectively. Thus if $t \in \mathbb{T}_i$, then $\mathbf{X}_i(t)$ belongs to the training dataset and has a known activity label $L_i(t)$; otherwise $L_i(t)$ is unknown and is to be estimated. Training sets contain continuous segments or blocks of time to include full examples of each movement type.

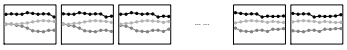
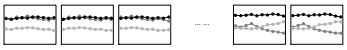
Next we define movelets as elements of time series that characterize movement in temporal windows with length H. More specifically, let

$$M_i(t) = \{\mathbf{X}_i(t), \mathbf{X}_i(t+1), \ldots, \mathbf{X}_i(t+H-1)\},$$

define the movelet of subject/visit i at time $t \in \{1, 2, \ldots, (T_i - H + 1)\}$. Note that movelets are made up of time series for all axes of the accelerometer output, and summarizes the pattern of acceleration recorded from time t to $t + H - 1$. The dimension of the movelet $M_i(t)$ is $3H$, because there are 3 concatenated time series, and contains all the accelerometry information for a window of movement of length $H/10$, because time is expressed in 10 Hz in our case. H is usually chosen so that a movelet $M_i(t)$ captures enough information to identify a movement and is not too long to contain more than one type of activity as well. Movelets $M_i(t)$ with $t \in \mathbb{T}$ are paired with their known activity labels and collected into activity-specific "chapters". Thus, we define a chapter $\mathcal{C}_a$ as a collection of movelets $\{M_i(t) : L_i(t) = \text{Act}_a\}$ that share a common label. An important characteristic of movelets is that they are overlapping moving windows; in fact $M_i(t)$ and $M_i(t+1)$ overlap everywhere, except at time t and $t + H$. This is an important characteristic when there is uncertainty on where the activity actually starts, because transitions between two activities can be unclear particularly for elderly subjects. This happens to be a serious problem even with the best in-lab human annotation. Allowing this sort of obscure period in our movelets may help us solve the problem. One chapter is constructed for each activity type; chapters are then combined to form a subject-visit specific "dictionary" of movelets and their labels. Dictionaries are distinct for subjects and visits to control for differences between the movement patterns for different subjects and to account for changes in the orientation of the accelerometer at different visits. This dictionary

Table 2.1: A subject-specific dictionary with with A chapters, one for each activity type. Each chapter consists of movelets, short overlapping segments of three-axis accelerometer data, which are illustrated in the far-right column of the table

Dictionary

Chapter	**Activity**	**Movelets**	
$\mathcal{C}_1$	Activity 1	$\{M_i(t) : L_i(t) = \text{Act}_1\}$	
$\mathcal{C}_2$	Activity 2	$\{M_i(t) : L_i(t) = \text{Act}_2\}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{C}_A$	Activity A	$\{M_i(t) : L_i(t) = \text{Act}_A\}$	

is used as a reference for movelets $M_i(t)$ with $t \in \mathbb{V}$. Table 2.1 displays an example of a subject-specific dictionary consisting of A chapters in total. Each chapter is constructed using the training set and is made up of movelets, the short components of three-axis accelerometer data. Usually for activities with well-defined beginning and endings(standing up from chair, etc) one full replicate is used to construct a chapter. For continuous activities(walking, sitting, etc.) we use a two-to-three-second segment to build the chapter.

The definitions of movelets, chapters, and dictionaries given above provide a useful analogy for our proposed classification method. Given unlabeled accelerometer data that has been decomposed into movelets, we use the dictionary as a reference by "looking up" an unlabeled movelet and finding its best match among known movelets. The label associated with the best match, which is the chapter title, is used to predict the unknown label. Matching, which is described below, quantifies the intuition that

movelets with similar visual appearances are likely to be components of the same larger movement.

2.2.2 Matching and labeling

Given an unlabeled movelet $M_i(t_0)$, we predict the label $L_i(t_0)$ first by matching $M_i(t_0)$ to a chapter in the dictionary described above. To be more specific, the closest match for movelet $M_i(t_0)$ in the dictionary is $M_i(t')$, where

$$t' = \underset{t \in \mathbb{T}}{\operatorname{argmin}} \left[D\{M_i(t), M_i(t_0)\} \right].$$

The distance function $D(\cdot, \cdot)$ is

$$D[M_i(t_1), M_i(t_2)] = \frac{1}{3} \sum_{p=1}^{3} \sqrt{\sum_{h=1}^{H} [X_{ip}(t_1 - 1 + h) - X_{ip}(t_2 - 1 + h)]^2}. \tag{2.1}$$

Thus, distance between movelets averages the difference taken over all acceleration axes. Based on this match, an estimate for the unknown label is $L_i^*(t_0) = L_i(t')$; that is, we take the label associated with the best dictionary match and use it to estimate the unknown label. Figure 2.2 gives a schematic of the matching process, in which an unlabeled movelet $M_i(t^*)$ is compared to a dictionary with 4 chapters. The distance between $M_i(t^*)$ and all reference movelets is calculated using the distance function (2.1). After $M_i(t^*)$ is compared to all reference movelets in the dictionary it is matched to Chapter 2, because movelet $M_i(t')$ in Chapter 2 along with $M_i(t^*)$ provides the smallest distance.

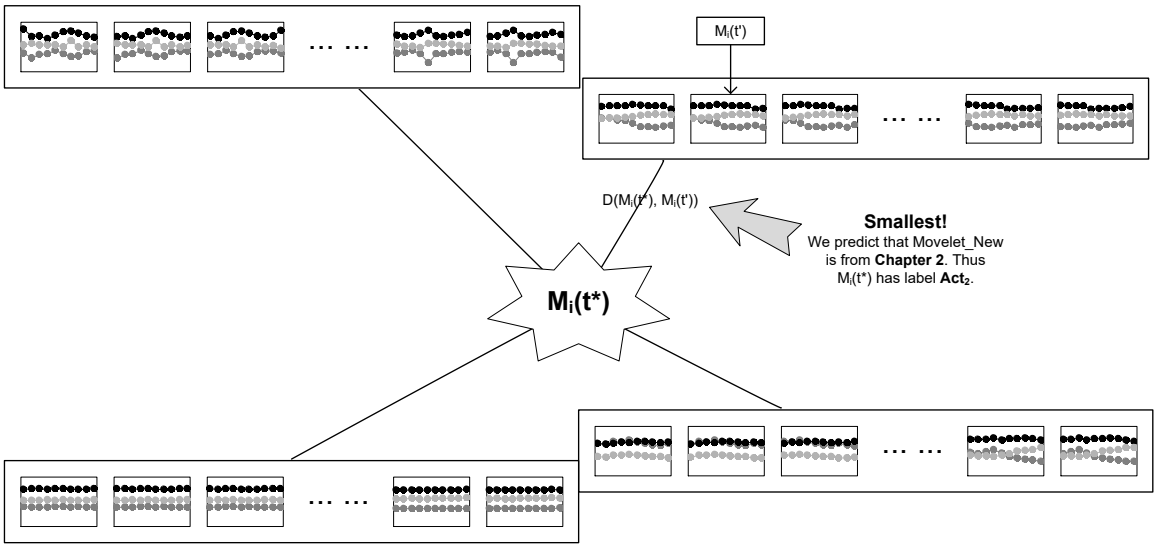


Figure 2.2: A display of matching an unlabeled movelet $M_i(t^*)$ to 4 chapters in the dictionary. Points in each chapter represent labeled movelets corresponding to the activity associated with this chapter. The distance between the unlabeled $M_i(t^*)$ and each chapter is given by the minimum distance between $M_i(t^*)$ and the movelets in each chapter. After $M_i(t^*)$ is compared to all reference movelets in the dictionary, it is matched to Chapter 2 which provides the smallest distance among all the 4 chapters.

After preliminary labels $L_i^*(t)$, $t \in \mathbb{V}$, are generated using the matching step, a majority voting procedure is used to select final estimated labels $\hat{L}_i(t)$. Each element of $\{L_i^*(t), L_i^*(t+1), \ldots, L_i^*(t+H)\}$ $(t \in [0, T_i - H])$ is considered a single vote, and the activity with the most votes in this set is the estimate $\hat{L}_i(t)$. An advantage of this procedure is that it smooths the predicted labels $\hat{L}_i(t)$ by taking into account the fact that movements are continuous, meaning that neighboring movelets contain information about the current activity. Additionally, because movelets decompose movements into their constituent parts, the matching applies even when the duration of movements is variable. For instance, two replicates of sitting from the standing position may take different amounts of time, but will have similar movelet signatures.

2.2.3 Movement fingerprints and lazy movelets

To increase the accuracy of our dictionary-based classification method and decrease the computational burden of the looking-up process, each chapter must be carefully constructed to include useful information while excluding redundant or less useful movelets. With this in mind, chapters that were built in the manner described above can be fine-tuned using the identification of what we will label "fingerprint" and "lazy" movelets.

First, each chapter must include the signature movelets of the corresponding activity. We refer to these defining movelets as "fingerprints" because they provide excellent prediction of a specific activity related to the chapter. Fingerprints are thus the characteristic acceleration time series associated with a movement, and are most often used when matching new movelets of the same activity. Second, unnecessary or redundant information should be removed from the chapter. For example, a chapter built on several seconds of walking will include many near-identical movelets due to the periodic nature of the activity. Further, there often exist "lazy" movelets which, contrary to fingerprints, are not commonly matched to and do not usefully identify the activity; rather than aiding prediction, these can be falsely matched to by movelets of other activities. Both redundant and lazy movelets can be excluded from a chapter to increase computational performance and reduce the number of errors. Finally, some movements share very similar movelets. These "ambiguous" movelets can lead to misclassification due to very close matches in multiple chapters. In this

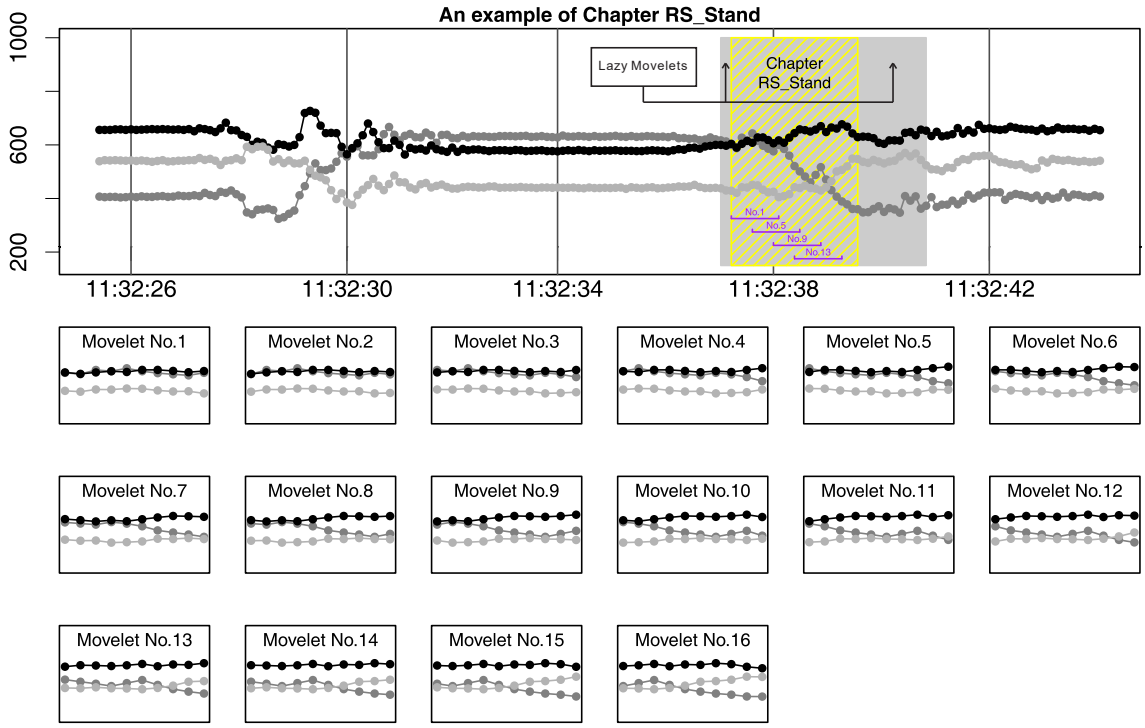


Figure 2.3: The chapter "Standing from Lying", which consists of 16 movelets. In dark grey is the section of the acceleration data used to construct the chapter; in light grey are time points with the same activity label, but that are excluded from the chapter as "lazy" movelets.

situation, an ambiguous movelet can be removed from one chapter so that matches will be made to the remaining movelet; the choice of which movelet to retain will depend on the relative importance of correctly classifying the two movements. The selection of fingerprint and lazy movelets was done independently of performance on the test set.

As an example of both fingerprints and lazy movelets, Figure 2.3 displays the chapter for "Standing from Lying" from a movelet dictionary. We used only the yellow-line-shaded region to construct the chapter, despite the fact that the areas

shown in light gray are also labeled by a human observer as "Standing from Lying". The fingerprint of this activity is the pattern that the mid gray time series goes down while the green one goes up. The movelets in the light gray bands (not shaded by yellow lines) are lazy movelets, and do not distinguish this activity from others. We removed the lazy movelets from the annotated time period and built the library conservatively to make the chapter a more useful reference for future unlabeled activities.

2.2.4 Summary

Movelet-based analysis of accelerometer data is built on the intuition that movements with similar acceleration patterns at the elemental level are likely to be generated by the same activity. Using this idea, we decompose movements into overlapping segments and construct reference chapters and dictionaries; given unlabeled time series, we match to the reference and use the best match to predict the unknown activity type. Movement fingerprints are identified to strengthen the construction of chapters and to aid in the basic understanding of movements, while lazy movelets are eliminated to reduce classification error and computation time. The result is a conceptually clear method for activity prediction that is computationally feasible and scalable to large datasets.

2.3 Application to LIFEmeter data

We now apply our methods to data from two subjects, each with two visits. Data were collected in the development of the LIFEmeter multi-sensor device, intended to assess physical function in large-scale observational studies. The subjects were community dwelling older adult participants in the LIFEmeter study, ages 65 and older who had no history of cognitive dysfunction, lived in the Baltimore area, and were capable of walking across a small room unassisted. They were observed in a clinical setting, and performed physical activities that are common in daily living. The following activities were selected as important in understanding physical function in real-world setting: walking, standing from sitting, standing from lying, sitting from standing, and lying from standing. Three sedentary states (standing, sitting, and lying) were also annotated. Table 2.2 lists all activities observed and provides abbreviations that will be used through the remainder of this section.

An observer annotated the time points at which an activity was started and completed, providing activity labels $L_i^{obs}(t)$. Annotations were imperfect due to early or late start and stop points, to rounding times to the nearest second, and to misalignment. Obvious errors in the observed labels were detected and corrected through comparison with the accelerometer output to create labels used to construct movelet dictionaries and assess the predictive performance of our algorithm.

Table 2.2: A list of activities of interest, with abbreviates used in remaining Figures and text

Activity List

Activity	**Alias**
Rest (Stand)	Standing
Rest (Sit)	Sitting
Rest (Lie)	Lying
Standing from Chair	CS_Stand
Sitting Down from Standing	CS_Sit
Lying Down from Standing	RS_Lie
Standing from Lying	RS_Stand
Walking	Walk

2.3.1 Constructing the dictionary

Following the method described in Section 2.2, we build a dictionary with 8 chapters of activities for each subject and visit. First, we partitioned the accelerometer data into training and validation sets $\mathbb{T}_i$ and $\mathbb{V}_i$. Using the training set, we decompose movements into movelets and organize by activity type. Our choice of H is 10, based on the 10Hz sample rate of the device used in our data collection. This is because each 1-second movelet contains just enough information to identify a movement, and is not so long that it restricts the matching of an unknown activity. We also tried other choices of H between 10 and 15 which did not give substantially different re-

sults. We therefore conclude that, in general, the methods is robust to the choice of H within a reasonable range (in our case around 10). For activities with well-defined beginnings and endings, such as "CS_Stand" and "CS_Sit", we use the first replicate as training data and reserve the remaining replicates as testing data. Chapters for these activities contain between 5 and 30 movelets each, depending on the duration of the activity. For continuous movements that lack well-defined beginnings and endings, such as "Walk" or "standing", we extract segments lasting 2 to 3 seconds that are clearly labeled with a particular activity to build the corresponding chapter. This is done to prevent chapters from becoming too large, and, since these activities are periodic, to prevent redundant information from being included in the reference.

2.3.2 Initial results

After constructing dictionaries for each subject and each visit using the training data, we predict activity labels $\hat{L}_i(t)$ for $s \in \mathbb{V}_i$ by matching movelets to the reference and implementing the majority voting step. Figure 2.4 details this analysis. For the accelerometer data displayed in Figure 2.1 (one segment of walking and two replicates of lie-rest-stand), the lower panel of Figure 2.4 shows the minimum distance between each unlabeled movelet and all movelets contained in the reference chapters as a collection of distance curves. The preliminary labels $L_i^*(t)$ are taken to be the chapter title with smallest distance. Next, the prediction $\hat{L}_i(t)$ is determined via a majority vote in which each element of $\{L_i^*(t), L_i^*(t+1), \ldots, L_i^*(t+H)\}$

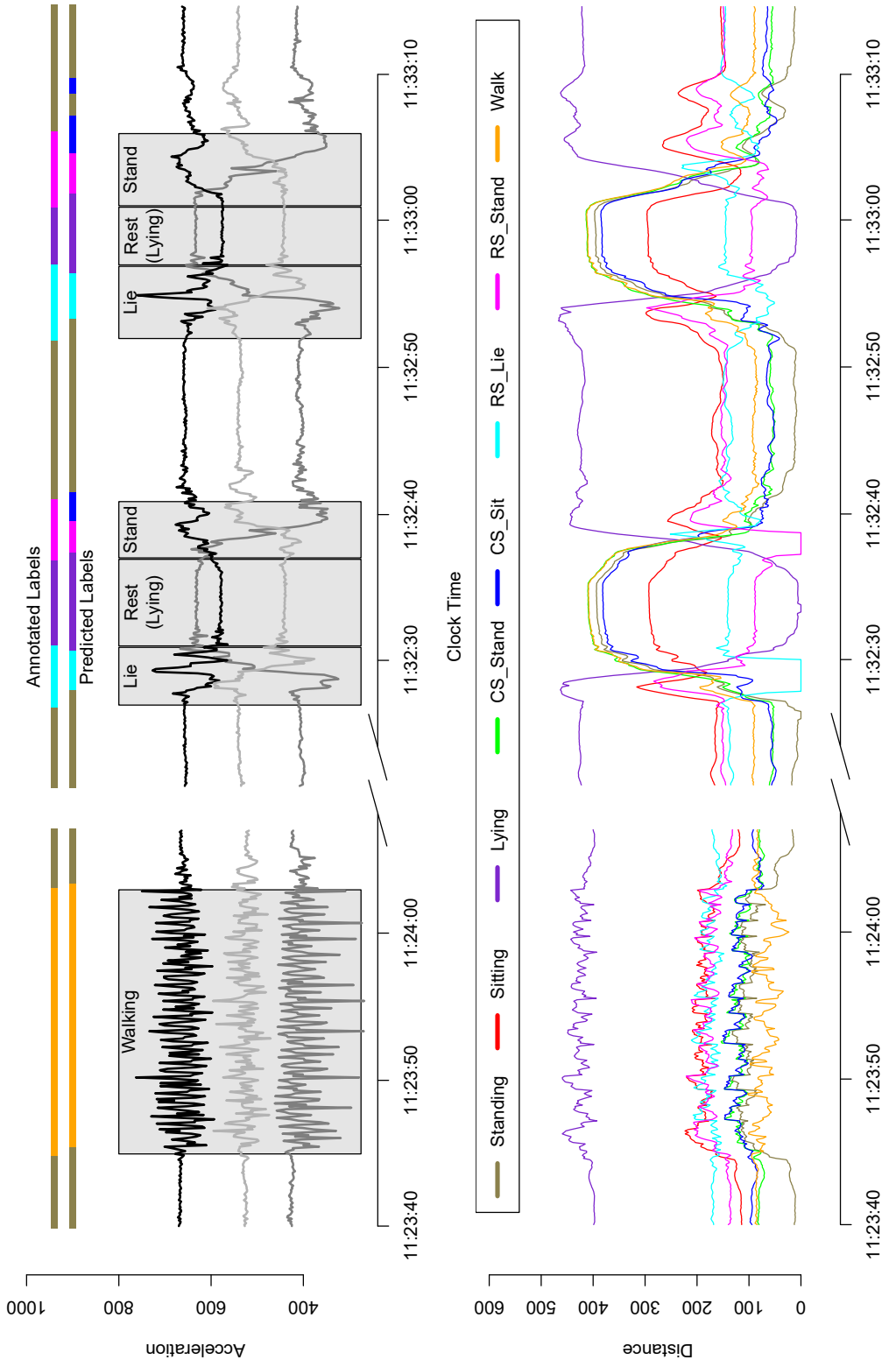


Figure 2.4: Observer-defined annotations and predictions for two segments of accelerometer data with several activity types. Curves giving the smallest distance between movelets and each chapter are displayed.

($t \in [0, T_i - H]$) is considered a single vote. At the top of Figure 2.4 are the observer-annotated (top colored bar) and predicted labels (bottom color bar) that accompany the accelerometer data. A comparison of the annotations and predictions indicates generally high agreement between these time series. In particular, there is broad overlap between the prediction and annotation of walking and resting periods as well as the location of the shorter activities lying and standing. Moreover, there is generally reasonable separation between the distance curve corresponding the the correct chapter and the remaining chapters, indicating the ability of the movelet-based analysis to distinguish between activity types. In two regions, the distance curves are zero – these depict the first replicate of the "Lie from Stand" and "Stand from Lie" activities, and were used to construct their respective activity chapters. Isolated misclassifications in the preliminary labels, such as those that take place in the middle of walking period, are in effect smoothed by the majority-voting step which prevents single activity labels from disagreeing with its neighbors.

On the other hand, as shown in the right segment of Figure 2.4, the annotated labels for the shorter activities have much longer time durations than the predicted intervals. This is most likely due to a combination of early and late stop points in the annotations and time spent transitioning between activities. For example, when a subject is asked to sit from a standing position, there is a brief pause as the new movement is begun; similarly, when rising to a standing position, there is a short period

of stabilization as the movement is completed. The extent to which these transitions will appear in real-world data, rather than in a controlled setting, is unclear. In these periods, the "true activity" is not clearly defined but the annotations are seen to be conservative in starting and stopping short activities, whereas the predictions extend neighboring (well-predicted) resting periods. This contrast can negatively affect the apparent prediction accuracy, although many of the activities are correctly identified.

Let V_i^a be the amount of time spent performing activity a (measured by $L_i^{obs}(t)$) and $\hat{V}_i^a$ be the predicted amount of time spent performing activity a. For each subject and visit, in Table 2.3, we report $\hat{V}_i^{a'}/V_i^a$ for all activities a, a'.

Table 2.3 reinforces the observations from Figure 2.4 that long continuous activities, like resting and walking, are better predicted than short activities, like standing from a chair. In fact, with the exception of subject 1 at visit 1, all resting states are accurately predicted more that 99% of the time, and walking is accurately predicted between 68% and 80% of the time. However, short activities seem to be fairly poorly predicted, and are often mistaken for one of the resting states. Again, this apparent shortcoming stems from two major factors: *i*. these activities are undertaken for very short periods, so even minor misclassification can greatly impact results, and more importantly *ii*. the observer-provided annotations for these short activities are inaccurate.

Table 2.3: Comparison of observer-annotated labels $L_i^{obs}(s)$ and the predicted labels $\hat{L}_i(s)$, expressed as the proportion of the predicted time spent engaged in an activity and the time spent engaged in the activity according to the annotated activity labels

Subject 1 Visit 1

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	78.3%	0	0	21.7%	0	0	0	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	4.3%	30.5%	0	55.3%	0	0	6.4%	3.5%
CS_Sit	11.9%	37.3%	0	23.1%	27.6%	0	0	0
RS_Lie	0	0	22.9%	49.4%	0	27.7%	0	0
RS_Stand	0	0	21.3%	59.8%	0	0	18.9%	0
Walk	9.6%	0	0	10.4%	0.5%	0	0	79.5%

Subject 1 Visit 2

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	96.8%	0	0	0	2.7%	0	0	0.4%
Sitting	0	99.9%	0	0	0.1%	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	25.6%	25.6%	0	40.2%	8.5%	0	0	0
CS_Sit	25.7%	12.8%	0	0	57.8%	3.7%	0	0
RS_Lie	40.3%	0	14.9%	0	0	44.8%	0	0
RS_Stand	0	0	21.1%	2.8%	39.4%	0	36.6%	0
Walk	18.8%	0	0	0.2%	0.6%	1.1%	0	79.3%

Subject 2 Visit 1

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	99.9%	0	0	0	0.1%	0	0	0
Sitting	0.7%	99.2%	0	0.1%	0.1%	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	10.9%	16.4%	0	57.3%	0	0	10.0%	5.5%
CS_Sit	11.1%	44.4%	0	4.6%	34.6%	5.2%	0	0
RS_Lie	10.6%	0	50.6%	0	0	27.1%	11.8%	0
RS_Stand	36.8%	0	24.6%	0	0	0	38.6%	0
Walk	22.1%	0	0	0.3%	0	0.2%	1.0%	76.4%

Subject 2 Visit 2

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	100%	0	0	0	0	0	0	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	7.9%	40.4%	0	46.1%	0	0	0	5.6%
CS_Sit	33.3%	20.6%	0	9.8%	35.3%	0	0	1.0%
RS_Lie	42.6%	0	31.5%	0	0	25.9%	0	0
RS_Stand	34.4%	0	34.4%	21.3%	0	0	9.8%	0
Walk	31.3%	0	0	0.1%	0	0	0	68.6%

2.3.3 Refined results

A comparison of our initial predictions, the observer defined annotations and the raw accelerometer data indicate that a gold standard for $L_i(t)$, the true activity labels associated with acceleration data, is not given by the observer's annotations $L_i^{obs}(t)$. Thus, we next create a "combined observer" to define activity labels $L_i^{com}(t)$ by synthesizing information from the observer annotations and raw accelerometer output. Primarily, this resulted in designating times between two distinct activities as "transition times", rather than misleadingly assigning these periods to one or the other activity. The new activity labels are shown in Figure 2.5, and a comparison of labels $L_i^{com}(t)$ and predictions $\hat{L}_i(t)$ is given in Table 2.4. All the tables demonstrate the large improvements in prediction accuracy that arise from improvements in the standard used to define true activity labels. We contend that these findings indicate that: 1) accurate labeling is crucial to prediction algorithm training; 2) a large source of prediction inaccuracies can reliably be traced to human labeling; and 3) prediction accuracy results reported in the literature are hard to compare because data use different labeling protocols.

The construction of the combined observer also illustrates the feedback from the movelet-based prediction algorithm to the annotations. Periods that were largely misclassified using $L_i^{obs}(t)$ as a reference, and that were labeled as "transitions" in $L_i^{com}(t)$, are periods where the distance between an unlabeled movelet and those in the reference dictionary is large. Thus, movelets that don't match well to any

Table 2.4: Table of prediction agreement for both subjects and both visits, using the combined observer labels.

Subject 1 Visit 1

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	85.8%	0	0	14.2%	0	0	0	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	0	1.4%	0	93.0%	0	0	5.6%	0
CS_Sit	4.2%	0	0	41.7%	54.2%	0	0	0
RS_Lie	0	0	0	0	0	100%	0	0
RS_Stand	0	0	22.6%		0	0	77.4%	0
Walk	0	0	0	3.7%	0	0	0	96.3%

Subject 1 Visit 2

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	98.8%	0	0	0	1.2%	0	0	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	7.5%	0	0	75.0%	17.5%	0	0	0
CS_Sit	4.2%	1.4%	0	0	88.8%	5.6%	0	0
RS_Lie	0	0	16.7%	0	0	83.3%	0	0
RS_Stand	0	0	0	8.0%	0	0	92.0%	0
Walk	13.2%	0	0	0	0.4%	1.1%	0	85.3%

Subject 2 Visit 1

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	99.8%	0	0	0	0	0	0.2%	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	0	0	0	96.9%	0	0	3.1%	0
CS_Sit	0	20.0%	0	7.8%	60.0%	12.2%	0	0
RS_Lie	0	0	0	0	0	100%	0	0
RS_Stand	0	0	0	0	0	0	100%	0
Walk	9.0%	0	0	0	0	0	1.1%	89.9%

Subject 2 Visit 2

	Prediction							
Truth	Standing	Sitting	Lying	CS_Stand	CS_Sit	RS_Lie	RS_Stand	Walk
Standing	100%	0	0	0	0	0	0	0
Sitting	0	100%	0	0	0	0	0	0
Lying	0	0	100%	0	0	0	0	0
CS_Stand	0	0	0	100%	0	0	0	0
CS_Sit	6.1%	0	0	20.4%	73.5%	0	0	0
RS_Lie	0	0	0	0	0	100%	0	0
RS_Stand	0	0	0	68.4%	0	0	31.6%	0
Walk	13.7%	0	0	0	0	0	0	86.3%

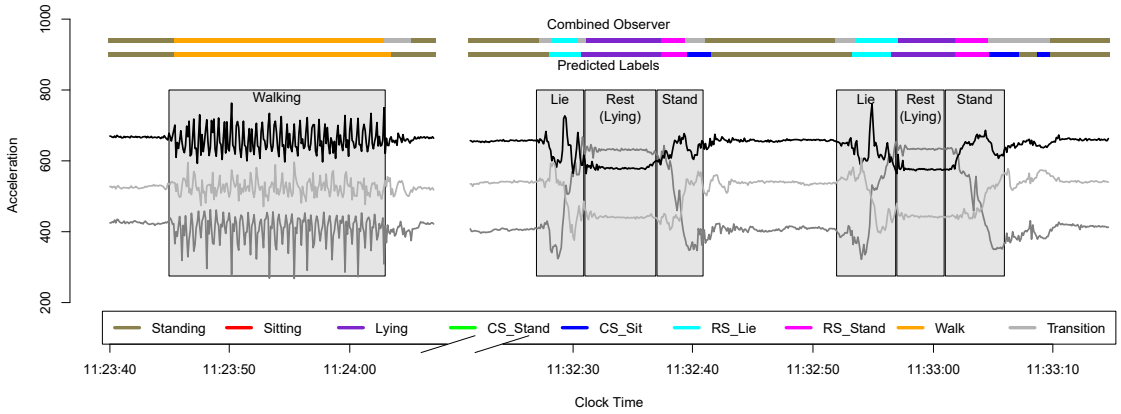


Figure 2.5: Comparison of "combined observer" annotations, based on observed-defined annotations and an inspection of the raw accelerometer data, and predicted labels.

known reference can be quickly identified. In observational studies, this facilitates the recognition of movements that are not included in any dictionary or are otherwise abnormal.

2.4 Discussion

Understanding physical activity is a key component in public health studies of subject function. However, standard measures of physical function such as activities of daily living questionnaires are subject to substantial measurement error. Emerging accelerometer technologies allow the collection of real-time, real-world activity data and may alleviate many of the issues with retrospective self-report data collection.

In this paper we propose a method for activity classification built around the "movelet" as a basic element of movements. Using movelets with known activities,

we construct reference chapters and dictionaries; given an unlabeled movelet, we find its closest match in the reference and use the match's label as a basis for prediction. Thus, our method is built on the intuition that movements with similar component acceleration patterns are likely to be generated by the same activity. This allows the method, and the matches it provides, to be quickly evaluated based on visual inspection of the accelerometer time series. Moreover, the extension to large data sets in which subjects are observed for hours or days is direct, because activity prediction is local in time. Finally, our method accurately predicts short activities, such as taking a few steps, as well as relatively rare and low-frequency movements such as rising from a chair.

Several directions exist for improving the movelet-based method. Focusing on the predictions for a single subject, transition models could naturally encode information about the order of movements and the likelihood of switching between them. Similarly, smoothing the distance functions (shown in Figures 2.4 and 2.5) would allow neighboring time points to influence the prediction at the current time. In our analysis, the movelet lengths were chosen to be 1 second; the sensitivity of predictions to this choice should be examined. Augmenting dictionaries to include objects other than movelets, for instance by adding measures of mean and variation, or to include sources of data other than the accelerometer, such as recorded speech or location information from a GPS device, could improve predictions. Our current method relies on models trained on each individual, and the solution of this issue is under

exploration. A statistical technique is being developed to normalize the orientation of the devices across subjects, in order to enable us to perform prediction using models trained by other subjects. This will also increase our understanding of heterogeneity in acceleration patterns between and within subjects. For instance, constructing a multi-subject dictionary would necessitate an understanding of movement fingerprints across several subjects.

Our results and methods suggest three improvements that could help the deployment of this technology to large epidemiological studies. First, there is an increasing need to minimize the effect of changes in accelerometer orientation that can occur during normal movements; this can perhaps be addressed by taking advantage on gyroscopic capacities in the SHIMMER^{TM}device. This would facilitate interpretation of the accelerometry data, especially in realistic scenarios where people wear these devices for extended periods of time, and also might allow the construction of dictionaries for use in populations. Second, the study could be more accurate if a human observer goes to the home of the participants, explains the setting up, carefully instructs the placement of the device and conducts a short testing period using a known sequence of common activities whose duration and type is carefully annotated. This would also resolve the problem of requiring subject-specific training of prediction algorithms, which was mentioned previously. It would also place a smaller burden on the participants. Finally, replication and calibration pre-studies should be conducted to ensure that prediction algorithms perform well on new subject or visit data.

The ability of technological solutions to improve the prediction of activity from accelerometer output is currently being evaluated. In the next phase of data collection, gyroscopic information will be used to normalize data to a constant vertical orientation. This may reduce the sensitivity of the movelet approach to rotations of the device that naturally occur as it is worn, and could also increase the comparability of movelets across subjects. Complementary improvements in the data collection via updated technology and in the activity prediction through refinements of the movelet approach will be needed to construct useful biomarkers of activity in large observational studies. The process of using and implementing new technologies in observational studies is a hard process filled with potential pitfalls. However, we find this challenge to be well worth undertaking by statisticians even before the beginning of the study in the design phase.

Chapter 3

Accelerometry metrics extracted from raw accelerometry data

3.1 Introduction

A commonly used outcome measure in aging research is the capacity to engage in activities of daily living (ADLs), or sentinel behaviors required to live independently. Conventional methods for measuring ADL include self-reported questionnaires or clinician ratings based on observed behavior [32,58], which have several limitations. First, self-reported activity may be subject to recall bias, which can be accentuated by the decline of cognition and memory. Second, these measurements provide only a snapshot of an individual's daily activity, while detailed minute by minute information is often missing. Thus, there is an increasing need for unbiased, detailed measurements

of sentinel behaviors that describe the underlying functional capacity of the individual and are not confounded by uncontrollable bias and measurement error. One possible solution is using wearable computing devices, which allow collection of real-time, densely sampled information on movement. These devices could serve as silent, unbiased, tireless and non-obtrusive recorders of actual human activity in a real-world context. However, translating information from high volume and complex data from wearable sensors into acceptable measurements can be done only by careful standardization and transformation to guarantee the validity and reproducibility of the measurements. In contrast, current measurements produced by software that accompanies these devices are expressed either in "activity counts" or Metabolic Equivalent of Task "MET" units. The most commonly used device, Actigraph, produces activity counts that, while formally defined [1], do not have a clear interpretation, and may not capture sufficient variability in older subjects. The MET units are even more problematic as they are based on population calibration equations that are severely biased at the subject level and in older adults. We propose data normalization and a set of novel, explicit, and interpretable metrics that can be used in medical and epidemiological studies.

Wearable sensors for different types of activity are deployed in an increasing number of studies [16, 19, 40, 74, 92]. Here we are concerned with accelerometers worn either in-field or in-lab by older, community-dwelling adults. Our focus is providing a set of simple activity measurements from ultra large, high density accelerometry data,

and providing evidence in support of their validity and usefulness in epidemiological studies. Some metrics have been proposed to extract and summarize information from accelerometry data, especially in sleep studies [14, 46, 53, 59]. All these studies have focused on gross summary statistics such as: ratios of sleep/wake duration, total sleep time, proportion of wake time after sleep onset, etc. Such summaries allow researchers to apply standard statistical methods, though they over-simplify the data. Another group of metrics focuses on reducing the raw three-dimensional accelerometry data to a one-dimensional proxy of subjects' activity along time. One such example is the "activity count", which converts the raw 3-axis acceleration measurements through various proprietary algorithms developed by accelerometer manufacturers. Based on activity counts, many studies categorize activities into different groups with sedentary, light, moderate, and vigorous intensity according to predefined thresholds, and obtain proportions of time spent doing each type of activity [54, 72, 84]. There are several limitations when using "activity counts". Indeed, the definition differs from manufacturer to manufacturer and may change within manufacturer when a new device is released. Thus, it is unclear whether activity counts are comparable [5, 38]. The interpretation of "activities count" is provided by some manufactures. For example, the Actigraph describes the process as "ActiGraph's original activity monitor, the 7164 model, utilized a mechanical lever capable of measuring the change in acceleration with respect to time (g/sec, where g is gravity or 9.806 m/s2). To suppress unwanted motion and enhance human activity, the acceleration signal was passed

through an analog band-pass filter, the output of which yields a dynamic range of 4.26g/sec ($+/-2.13$g/sec) at 0.75Hz (center frequency of the filter). Using a sample rate of 10 samples-per-second, this filtered signal was then digitized into 256 distinct levels by an 8-bit solid-state analog-to-digital converter, producing $4.26g$/sec per 256 levels or 0.01664 g/sec/count (each level is considered 1 count). When each filtered sample is multiplied by the sample window of 0.1sec, a resolution of 0.001664g/count is achieved." While this is an excellent technical description, it leaves many questions unanswered. First, it is unclear exactly what are the formulas and whether they are applied to each axis separately or combined. Second, the transition between the quasi-continuous signal and the number of g's in one second is not defined; this is a function that reduces 30 numbers (tri-axial at 10Hz sampling) into one number. Third, methods fundamentally depend on many software parameters as well as on the sensitivity of the chip. Small changes in thresholds, sampling rates, chip sensitivity, or number of count levels can lead to dramatic batch effects within and, especially, between manufacturers. Fourth, only in rare occasions are devices validated in real data or using replication. Fifth, the interpretation of a count is, probably, "some one-epoch summary of the acceleration that is between 0.01664 and 0.03328 g/sec", whose utility in a large observational study remains to be debated. We conclude that "activity counts" are actually not counting activities or steps as their name implies; instead they are a proxy of the acceleration within a time interval. The missing piece is a paper like the one we are putting here forth; we are currently unaware

of any paper starting from raw data and building explicitly either "activity counts" or other explicit metrics. Our paper has the following goals: 1) propose an explicit data processing pipeline for high-dimensional accelerometer data; 2) present a transparent, interpretable and implementable set of metrics; 3) validate these metrics via visual inspection, replicated in-lab experiments, and association studies with health outcomes.

We used data from older adults who were fitted with a high-definition three axis accelerometer "Shimmer" [18, 63] and asked to perform standard activities in a laboratory under observation. Then, subjects were asked to wear those devices for 5 consecutive days during normal activity. To analyze the massive free-living data ($>$ 18 million observations per subject), we processed and summarized the data into several metrics that are intuitive and reproducible. Special attention was given to normalization to ensure comparability of measurements across subjects and visits. We investigated the validity and reproducibility of these new measurements and their association with self-reported health status.

3.2 Data collection

3.2.1 Study population

Community-dwelling men and women were recruited from an ongoing cohort study on the multilevel determinants of cognitive function in older adults, The Baltimore

Memory Study (BMS, AG19604, Brian Schwartz, PI). For the LIFEmeter substudy, 125 older adult subjects were recruited and enrolled after BMS visit 4 or 5. The purpose of the LIFEmeter substudy (AG027481, Thomas A. Glass, PI) was to develop and test a sensor platform for capturing enacted function in older adults. Enrolled subjects were brought into a lab setting, given an interview, and asked to perform a series of standard activities under observation wearing a waist mounted pouch containing several sensors. Next, subjects were asked to wear the LIFEmeter array during waking hours for 3 to 5 consecutive days, removing the device during showering, swimming and sleeping.

3.2.2 Data description

Our data are generated using Shimmer^{TM}Unit by Shimmer Research [18], mounted on subjects' waist. The device uses a standard tri-axial accelerometer chip found in many cell phones and other devices (Freescale MMA7361) and records acceleration in three mutually orthogonal directions with a sample rate of 10Hz. The output consists of 3 voltage time series, which are proxy measures of acceleration. The time series exhibit complex variability in overall level, amplitude, frequency, correlation, and patterns along the time course of different activities.

Figure 3.1 displays two 2.5-minute data segments representing the raw 3-axis accelerometer data from two subjects, labeled 3208 and 3056. Many studies [7, 13, 19, 51, 73, 92] found that lack of accelerometer motion, which is a rough proxy for actual

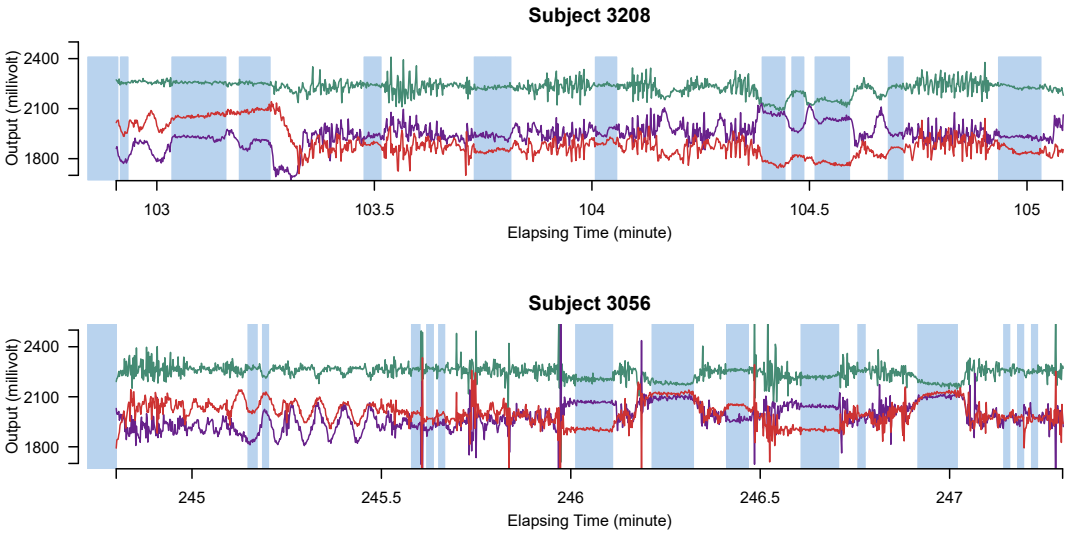


Figure 3.1: The raw 3-axis accelerometry data for Subject 3208 and Subject 3056 with Active v.s. Inactive prediction results. Each axis is illustrated in a different color. The x-axis stands for the time (in minute) from the first observation and the y-axis shows values of the raw output from the accelerometer. The inactive time periods (according to our algorithm which is also described in this section) are shaded in light slate gray.

human activity, is characterized by low variation around stable constants for *each* of the three time series. Using a simple method that will be described in this paper we have estimated periods of inactivity and shaded them in blue. An inspection of the time series for subject 3208 (upper panel in Figure 3.1) indicates that there are many periods of inactivity, each with a different length. Moreover, the accelerometer seems to be sensitive to different types of inactivity. Indeed, compare the blue block starting immediately after minute 103 with the one starting immediately after minute 104.5. There is low variability in both blocks, but the times series colored in red and purple have switched their mean levels. This probably indicates that the person is resting

in different postures (e.g. on a chair versus standing). Areas that were estimated to be active display a wide range of variation both in terms of patterns and amplitude of the signal. Inspecting such short time series is not dissimilar to listening to a new language, where we hear obvious patterns without having a clue about *what* is being said. However, it is quite easy to know *when* the person is not talking. A similar principle will be applied to identify periods of inactivity, by predicting areas of low variability above background.

We start by introducing notation. The observed data are a collection of three time series representing proxies of acceleration in three orthogonal axes. Denote the data (sample rate $f = 10$Hz) by $\mathbf{X}_i(t) = \{X_{i1}(t), X_{i2}(t), X_{i3}(t)\}^T$, $t = 1, 2, ..., T_i$, where T_i is the length of the accelerometer time series for Subject i. In this paper we used field data from 34 subjects and each subject was observed from 4 to 5 days. So $i = 1, 2, \ldots, 34$ and T_i is very large. For example, for a complete 5-day recording $T_i = 4,320,000$. Here we will be working directly with the raw voltage data, though our methods apply as well to data expressed in gravity units. Indeed, if $\mathbf{X}_i(t)$ is the collection of voltage time series then the gravitation data can be obtained by the formula [78] $\mathbf{g}_i(t) = R^{-1} \cdot K^{-1} \cdot [\mathbf{X}_i(t) - \mathbf{b}_i(t)]$. Here $\mathbf{g}_i(t) = \{g_{i1}(t), g_{i2}(t), g_{i3}(t)\}^T$ is the ratio of acceleration on the three axes to gravity, R^{-1} is an alignment matrix and K^{-1} is a diagonal matrix specifying the sensitivity of the sensor along each axis. In the remainder of the paper we focus on $\mathbf{X}_i(t)$ not on $\mathbf{g}_i(t)$. As our normalization procedure is a combination of several linear transformations of the raw signals, explicit

formulas can be obtained for $\mathbf{g}_i(t)$, as well.

We introduce a new time series, $L_i(t) \in \{0, 1\}$, as the time series of labels which describe whether the Subject i is estimated to be "active" or "non-active" at each time point t. Non-active time includes both the time when the subject was resting while wearing the device and the time when the subject took the device off. Thus, $L_i(t) = 1$ if Subject i is active at time point t and $L_i(t) = 0$ otherwise. $L_i(t)$ is observable either by study team members or from detailed diaries. In our study, we observe $L_i(t)$ only during in-lab sessions and not during the in-home data collection. We treat $L_i(\cdot)$'s as an unknown variable to be estimated during in-home monitoring. [7] introduced a method to classify accelerometry time series into active and non-active, which is essentially estimating the time-series of labels $L_i(t)$. This method applies a threshold on standard deviation in each one second interval. More specifically, for each time point t of Subject i, let $\sigma_i(t) = \{\sigma_{i1}(t)+\sigma_{i2}(t)+\sigma_{i3}(t)\}/3$, where $\sigma_{im}(t)$ $(m = 1, 2, 3)$ is the standard deviation of the mth axis of the acceleration time-series $X_{im}(t), X_{im}(t+1), \ldots, X_{im}(t+H-1)$ in a window of length H. Here we use a window of length $H = 10$, which corresponds to one second. We found this window size to work well in practice, as it reasonably corresponds to the temporal scale of human activity. More precisely, $\sigma_{im}(t) = [\sum_{h=0}^{H-1} \{X_{im}(t+h) - \frac{1}{H}\sum_{k=0}^{H-1} X_{im}(t+k)\}^2/H]^{1/2}$. We will not use notation that depends on the window size, as $H = 10$ is fixed throughout this paper. Also, we do not use hats to indicate that standard deviations are estimated, as we are primarily focused on prediction and algorithmic signal extraction and not

on inference.

Once the subject-time specific standard deviation, $\sigma_i(t)$, of the signal is computed, the activity label time series is estimated as $L_i(t) = 1$ if $\sigma_i(t) > C$ and 0 if $\sigma_i(t) \leq C$. Here C is a threshold value that does not depend on the subject and is estimated from the data. Here we investigate the impact of various choices of C and are especially interested in finding whether a common threshold is reasonable for all subjects. The threshold C will depend on the scale and type of output used in the analysis. As our data were collected in millivolts, C is also expressed in millivolts. If data $\mathbf{X}_i(t)$ were expressed in g's, we could also specify C in g's. We have tried threshold approaches for both type of signals and obtained indistinguishable results; this should not be surprising given the one-to-one and monotonic relationship between the millivolt and g scales.

To further investigate the threshold we calculated the standard deviation, $\sigma_i(t)$, for each of the 34 subjects in each 1 second interval. For each subject we produced a smooth histogram of standard deviation values collapsed over time. The top panel in Figure 3.2 displays these 34 density curves for all subjects, each in a different color. A feature of these curves is that they all have a high peak centered roughly around the same value (2-3 millivolts) with much of the mass concentrated between 1 and 7 millivolts. The reason is that people spend a lot of time resting. These plots suggest that a cut-off point of $C = 10$ millivolts could separate active from inactive periods in all subjects. For example, the blue shaded areas indicating inactive periods

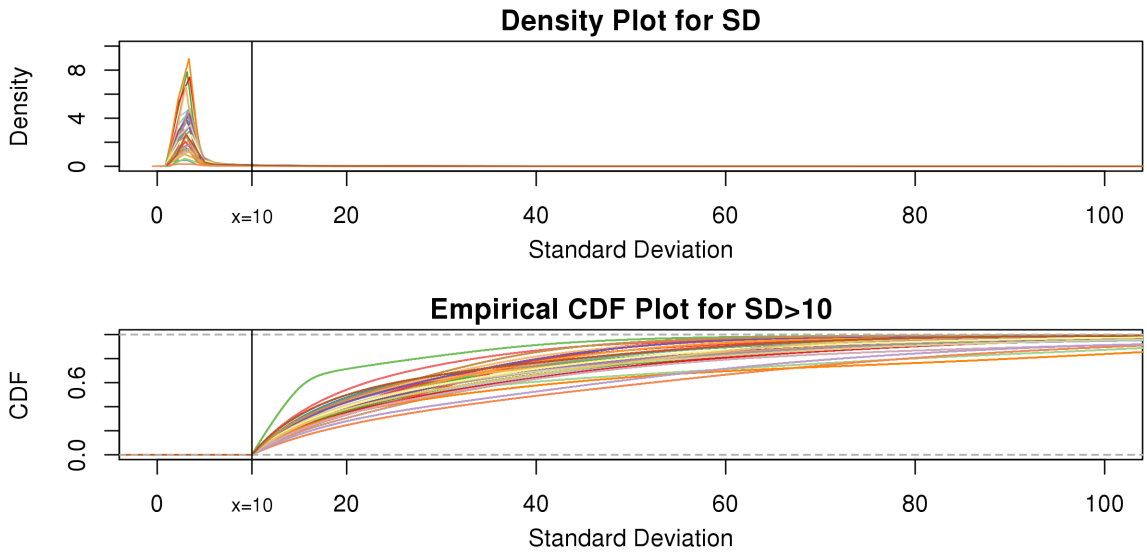


Figure 3.2: The density curves of standard deviations for all (34) subjects. The upper panel contains density curves of the original standard deviations and the lower panel contains density curves of standard deviations greater than $C = 10$. The vertical black lines in both panels are at $x = 10$.

in Figure 3.1 are obtained using this decision rule. We have checked several other thresholds between 8 and 15 millivolts and they provided similar results. The reason is that there are few activities that are visually identifiable and correspond to a standard deviation in the range $[10, 15]$ millivolts. While some ambiguity is likely to remain even after careful visual inspection of each time series, we conclude that $C = 10$ millivolts works well for our data set.

All histograms have long tails, which correspond to visually identifiable activities. As accelerometer time series are likely to be dominated by inactive periods, the top panel in Figure 3.2 does not display enough detail to understand subject-to-subject differences in activity intensity. Thus, the bottom panel displays the cumulative

distribution function for standard deviations above $C = 10$ millivolts. We focus on the green curve that is on top of the other curves. A value of 0.7 of the cumulative distribution function at 20 millivolts indicates that about 70% of standard deviation values that are higher than 10 millivolts are between 10 and 20 millivolts. This implies that this person has lower intensity movements compared to the other subjects. If a subject-specific curve is higher for a given subject it indicates that the first subject has lower level of activities across the range of observed activities. The fact that the curves do not seem to cross each other indicates a reasonable finding: if subjects tend to have fewer low intensity movements they also tend to have fewer high intensity movements. This is consistent with an elderly population, though it should not be surprising to find similar patterns in other age ranges. Once the active or non-active labels $L_i(t)$'s are estimated, there were many possibilities for estimating various types of metrics to describe the rest of the data. We start by dividing the entire recorded time period of Subject i into two sets of time points, T_i^A and T_i^I, corresponding to active and inactive time periods. Specifically, $\forall t \in T_i^A$, $L_i(t) = 1$ and $\forall t \in T_i^I$, $L_i(t) = 0$.

3.3 Accelerometer metrics definition

In this section, we introduce several metrics that were found to be sensible and feasible to compute. We denote by J_i the number of days when Subject i is observed,

while T_i is the total number of time points where the subject is observed. The number of days, J_i, varies between 3 and 5, whereas T_i is in the millions. We denote by t_{ij}^0 the time index for start of day j, which has a total of T_{ij} data points. For subject i on day j we propose to extract the following 5 dimensional vector of univariate signals labeled $D_{ij} = (T_{ij}, \text{TAM}_{ij}, \text{TAV}_{ij}, \text{AIM}_{ij}, \text{AIV}_{ij})$. Here T_{ij} is the length of time for the period estimated to be wake time and can depend on the particular day, j, because some days have shorter recording times or missing data. The variable "Time Active Mean", TAM_{ij}, represents the fraction of total time awake, T_{ij}, that was estimated to be active (non-rest). The variable "Time Active Variability", TAV_{ij}, represents the variability of the active/non-active process. The last two variables, "Activity Intensity Mean" and "Activity Intensity Variability", are similar to $\text{TAM}_{ij}, \text{TAV}_{ij}$, but focus on the actual intensity of movement (amplitude of signal) instead of the binary measurement active/non-active. We chose these five measurements only for simplicity, though they could be produced at much higher resolution, such as minute, hour, or time of day. A concern with reducing data sets of such complexity to a few summary measurements is whether this reduction is too aggressive. To alleviate this concern, we introduce two additional measurements, "Cumulative Relative Time Active (CRTA)" and "Cumulative Relative Activity Intensity (CRAI)". These measurements are calculated at every time point and preserve all the original information. In this paper we use them for visualization purposes, establish their characteristics, and defer their analysis to future publications.

3.3.1 Time active

After $L_i(t)$, the labels denoting whether Subject i is active at time t, were predicted for all subjects, they can be used to calculate each subject's "Time Active" within intervals of interest. To be specific, for Subject i, we first partition the whole time course T_i into non-overlapping windows of length W, with the total number of windows equal to $K = [T_i/W]$, where $[x]$ denotes the highest integer smaller than x. The window size can be anything, though here we focus on $W = 900$ seconds, which corresponds to 15 minutes. For a fixed time window of length W, we define the Time Active, $\text{TA}_i(k)$, for every $k = 1, \ldots, [T_i/W]$ as $\text{TA}_i(k) = \sum_{s=1}^{W} L_i\{W(k-1)+s\}/W$, which is the proportion of time declared active in the time window $[(k-1)W+1, kW]$ using the 10 millivolt threshold on the 1-second window standard deviation. This measure is useful because it: 1) is explicitly measuring the active time in a particular time window without combining it with the intensity of activity during the same period; 2) is easy to compute and reproduce given the original raw data; 3) is interpretable across subjects and devices; 4) expressed on a 0-1 scale; and 5) is not dependent on black-box software.

Figure 3.3 displays the original tri-axial acceleration 15-minute time series plots for Subject 3092 in panels 1 and 2. We display only 15 minutes of the raw accelerometer data, as showing the entire 5-day period would be daunting and quite useless for visual inspection. In contrast, the Time Active plot in the bottom panel provides a simple visualization tool for the entire duration of the study. In the bottom panel,

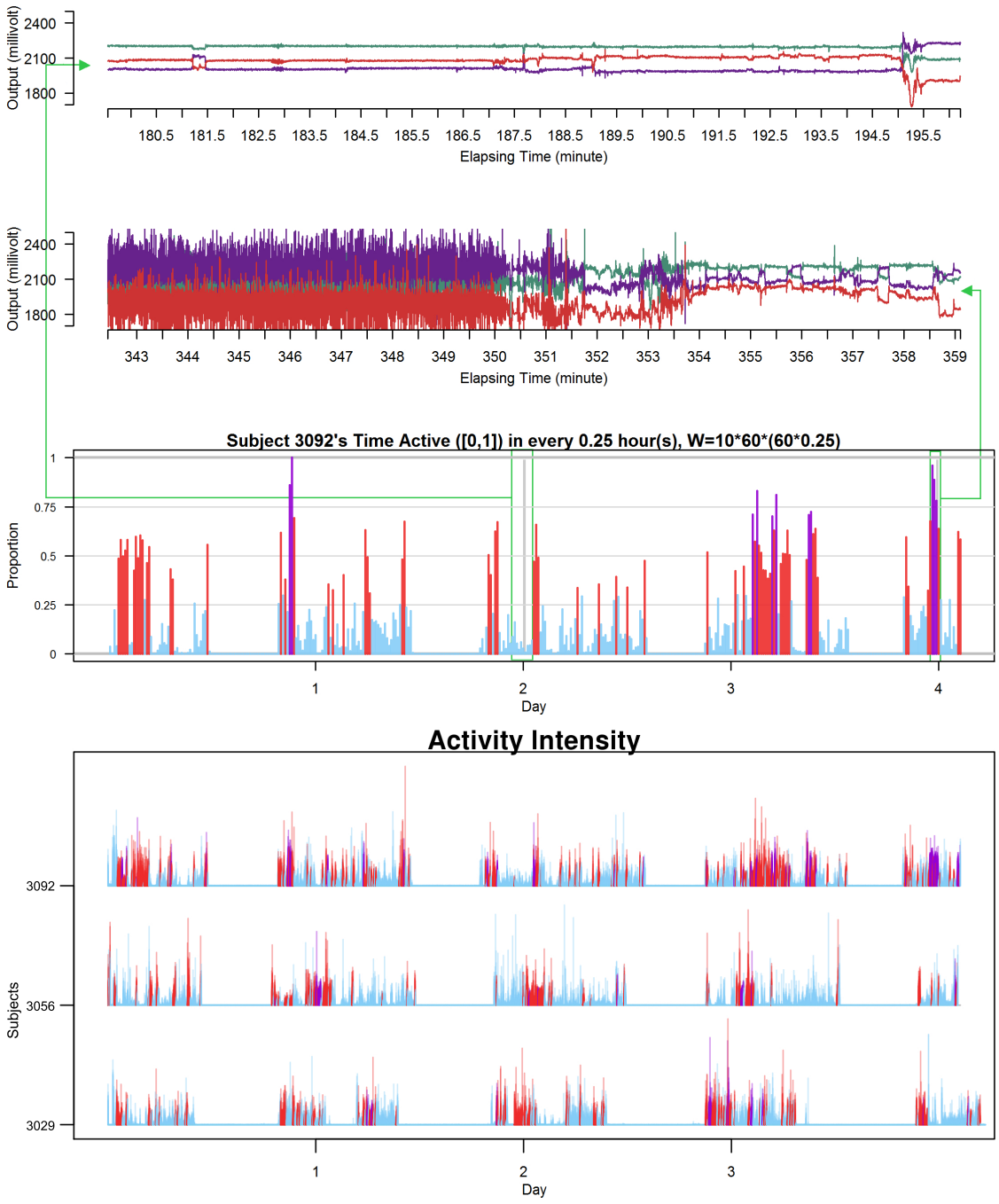


Figure 3.3: Two periods of raw data (panel 1 and 2), TA bars for Subject 3092 (panel 3). Each TA bar has a value between 0 to 1, and is colored light blue (TA ≤ 0.3), red ($0.3 <$ TA < 0.7) or purple (TA ≥ 0.7). Panel 1 and 2 correspond to the gray bars depicted in panel 3. Panel 4 displays 3-day AI for Subjects 3029, 3056 and 3092. AI bars are colored light blue, red or purple according to their TA values as in panel 3.

the Time Active (TA $\in [0, 1]$) bars of every 15-minute interval for the same subject are in light blue (TA ≤ 0.3), red ($0.3 <$ TA < 0.7) or purple (TA ≥ 0.7). Note that, for example, a blue bar means that the subject movement was distinguishable from inactivity according to the 10 millivolts threshold for up to $15 \times 0.3 = 4.5$ minutes out of the corresponding 15 minutes period. This plot corresponds to the 5 day period when Subject 3092 wore the device. As the raw and time active data are linked, one can always go back to a particular specific period for further visual inspection. TA has several long sections (i.e. between days) where it is zero, most likely during sleeping when the subject placed the device on a table. To better understand the data transformation, we placed two green boxes, each with a vertical gray bar, in the third panel. Each vertical gray bar represents a 15-minute period; the corresponding raw data are shown in the panels 1 and 2. Plots indicate that the data transformation is quite sensible. Indeed, the first framed time period of Subject 3092 has much lower time active values (0.01 v.s. 0.76) compared to the second framed time period. This difference can be easily observed by comparing the upper panel and mid panels in Figure 3.3.

3.3.1.1 Scalar summaries of time active

Time active mean is $\text{TAM}_{ij} = \sum_{s=1}^{T_{ij}} L_i(s + t_{ij}^0 - 1)/T_{ij}$, which is the average number of active periods, and $\text{TAV}_{ij} = \sqrt{\sum_{s=1}^{T_{ij}} \{L_i(s + t_{ij}^0 - 1) - \text{TAM}_{ij}\}^2 / T_{ij}}$, which is the standard deviation of the active periods for Subject i on day j. The two

measurements, TAM_{ij} and TAV_{ij} are complementary. Indeed, a subject with large TAM_{ij} and small TAV_{ij} would tend to have long periods of activity with few rests; a subject with small TAM_{ij} and large TAV_{ij} would tend to be less active but with short and sustained activity periods.

3.3.1.2 Cumulative relative time active

Using the time active, $\text{TA}_i(k)$, is not straightforward. Indeed, a quick inspection of the third panel in Figure 3.3 provides a reasonable summary, but leaves many questions unanswered: 1) how to handle the "spiky" nature of the data; 2) what to do about the de-synchronized behavior both within and between subjects; and 3) how to preserve the complex nature of the data without losing interpretability? To answer these questions we follow the idea introduced for displays of actigraphy data by [82]. We introduce the "Cumulative Relative Time Active" $\text{CRTA}_{ij}(t) = \sum_{s=1}^{t} L_i(s + t_{ij}^0 - 1)/T_{ij}$, which is the fraction of active periods up to time t of day j for Subject i out of the total time awake, T_{ij}. This approach provides a much smoother representation of the data while maintaining all the information. As the accelerometer is taken off during sleep, time can easily be partitioned into sleeping (non-wearing) and being awake (wearing).

Figure 3.4 (top panels) provides the CRTA for three different days for three subjects. Functions are displayed with respect to the proportion of time, $(s+t_{ij}^0-1)/T_{ij} \times 100\%$, from start of the day. Figure 3.4(b) indicates that Subject 3056's has similar

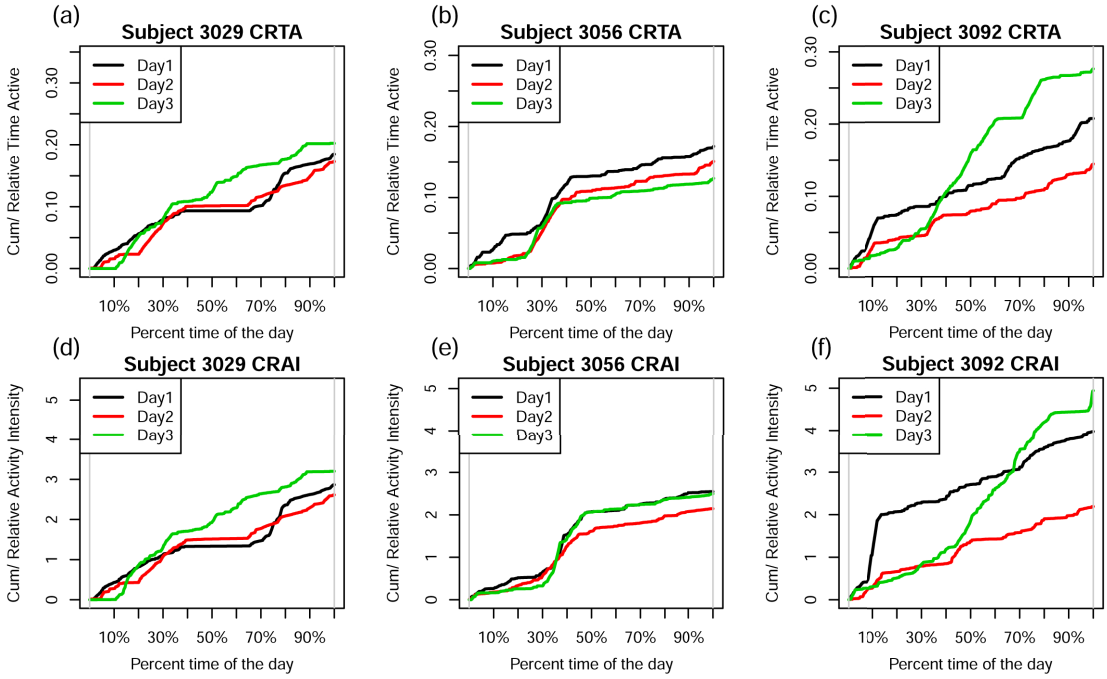


Figure 3.4: Top panels: Cumulative Relative Time Active (CRTA) for Subject 3029, 3056 and 3092. There are three curves in each plot, each representing one day. These curves are either compressed or stretched so that they display the CRTA in the scale of percentage time of the day, instead of actual time. Bottom panels: Cumulative Relative Activity Intensity (CRAI) for Subject 3029, 3056 and 3092. There are three curves in each plot, each representing one day. These curves are either compressed or stretched so that they display the CRAI in the scale of percentage time of the day, instead of actual time.

CRTA patterns for the three days. This subject is quite active in the middle of the day, which is indicated by the synchronized jumps in the day-specific curves around the 30%-50% section of the x-axis. Moreover, the black curve is higher suggesting that Subject 3056 spent more time being active on Day 1 than on either Day 2 or Day 3. Figure 3.4(c) suggests a different activity pattern for Subject 3092. On Day 1 and Day 2, the subject shared a similar CRTA pattern, but with different end points.

After a large jump around 10% of Day 1, the black curve remains roughly parallel to the red curve. This suggests that Subject 3092 had a short but active period after getting up on Day 1, but spent the rest of the day with an active/inactive pattern similar to that of Day 2. In contrast, on Day 3, the subject did not spend much time being active until the middle of the day, but then became very active for the rest of the day, leading to a high TA for that day. Such trends and differences seem obvious in Figure 3.4, but are hard to notice in Figure 3.3.

3.3.2 Activity intensity

Time active is a measure of *how long* the person was active without information about *how intense* the activity was. Herewe propose measurements that describe the entire spectrum of activity intensities. We estimate "Activity Intensity" as the standard deviation of the raw accelerometer signal relative to the standard deviation in the signal during non-wearing or rest. Thus, activity intensity will be expressed in sigma units, where sigma is the variation of the time series during non-wearing or rest. This approach has the potential to mitigate some of the inherent problems associated with accelerometer measurements. First, data will be normalized on a scale that can be interpreted in units of intensity of activity relative to the systematic noise (non-wearing time variability of the signal). This may reduce device- and day-specific systematic deviations in measurements. Second, measurements from the same type of accelerometer from different people and locations are more comparable and

have similar interpretation: observed variability *relative* to signal variability when the device is not perceptively moving. Third, the approach automatically mimics human information processing. Indeed, a human observer would naturally focus on areas of high, moderate and low variability; activity intensity quantifies this qualitative process.

We define $\sigma_i(t) = \{\sigma_{i1}(t), \sigma_{i2}(t), \sigma_{i3}(t)\}$ as the local standard deviation at time point t calculated in each one second interval ($H = 10$). We estimate the average standard deviation of each axis during rest as $\overline{\sigma}_{im} = \sum_{t \in \mathcal{T}_i} \sigma_{im}(t) I\{L_i(t) = 0\} / \sum_{t \in \mathcal{T}_i} I\{L_i(t) = 0\}$. The summation $t \in \mathcal{T}_i$ stands for each time point t in the rest period $\mathcal{T}_i$. The average standard deviation $\overline{\sigma}_i = \{\overline{\sigma}_{i1}(t), \overline{\sigma}_{i2}(t), \overline{\sigma}_{i3}(t)\}$ quantifies the device-specific variation when the device is either not worn or the person is at rest. We estimate $\overline{\sigma}_i$ using the periods when the subjects was sleeping or resting. Activity intensity (AI) is defined as $\text{AI}_i(t) = \max([\{\sigma_{i1}(t) - \overline{\sigma}_{i1}\}/\overline{\sigma}_{i1} + \{\sigma_{i2}(t) - \overline{\sigma}_{i2}\}/\overline{\sigma}_{i2} + \{\sigma_{i3}(t) - \overline{\sigma}_{i3}\}/\overline{\sigma}_{i3}]/3, 0)$. Thus, $\text{AI}_i(t) \in [0, +\infty)$ and is the difference between the current observed standard deviation $\sigma_i(t)$ and the average standard deviation $\overline{\sigma}_i$ during non-wearing/rest periods $\mathcal{T}_i$, relative to $\overline{\sigma}_i$. The truncation at zero is done to ensure that every standard deviation below the average standard deviation at non-wearing is set to zero and that there are no negative AI values. We used the average variability in non-wearing/rest periods as reference because it characterizes the systematic variability of the device.

AI is a complement to TA, which only focuses on time spent *while* active without

information about *how* active the subject is. For example, walking and running continuously for 15 minutes give the same TA (which is 1 because the subject is active during the whole period), but have completely different AI. Running has much higher levels of acceleration variation compared to walking, which is characterized by a larger AI for running. To illustrate this, the bottom panel in Figure 3.3 displays AI for three subjects (3029, 3056 and 3092). Each bar stands for AI in a 1 second interval and is colored by their TA values (TA $\leq$ 0.3: light blue; 0.3 $<$ TA $<$ 0.7: red; TA $\geq$ 0.7: purple). The AI and TA plots for Subject 3092 indicate that AI has a similar temporal pattern, though spikier and on a different scale. To better understand the complementarity of AI and TA, it is worth taking a closer look at the AI plot for Subject 3092. In the middle of day 3 there are two areas with purple bars (TA $\geq$ 0.7) among many red bars. In the TA plot they are obvious, whereas in the AI plot they are not. Thus, Subject 3092 was performing low to moderate intensity activities continuously during the purple periods, while in-between the subject performed a series of high intensity activities but with more rest periods.

3.3.2.1 Scalar summaries of activity intensity

Once AI is introduced, the AI mean is defined as $\text{AIM}_{ij} = \sum_{s=1}^{T_{ij}} \text{AI}_i(s + t_{ij}^0 - 1)/T_{ij}$, which is the average Activity Intensity for day j of Subject i. Similarly, the Activity Intensity Variability is $\text{AIV}_{ij} = \sqrt{\sum_{s=1}^{T_{ij}} \{\text{AI}_i(s + t_{ij}^0 - 1) - \text{AIM}_{ij}\}^2/T_{ij}}$, which is the standard deviation of AI. These measurements are similar to TA, though they focus

more on the levels of activity and less on whether or not the subject moved.

3.3.2.2 Cumulative Relative Activity Intensity

Similar to CRTA, we introduce $\text{CRAI}_{ij}(t) = \sum_{s=1}^{t} \text{AI}_i(s + t_{ij}^0 - 1)/T_{ij}$, which is the cumulative sum of AI of Subject i up to time t in day j. Recall that $\text{AI}_i(t)$ is a measure of how much larger is the variability of the accelerometer time series data in a time window centered at t relative to its variability at non-wearing time periods. Thus, $\text{AI}_i(t)$ is a proxy for the instantaneous intensity of human movement as measured at the hip by an accelerometer. Thus, $\text{CRAI}_{ij}(t)$ is a proxy measure of cumulative energy measured at the hip during movement up to time t of the day. To mitigate the effect of different lengths of day we divide this cumulative sum by T_{ij}. The bottom panels in Figure 3.4 display $\text{CRAI}_{ij}(t)$ for the same three subjects shown in the top panels. However, CRAI does provide something different. For example, in Figure 3.4(e), Subject 3056 had almost the same pattern of CRAI on Day 1 and Day 3. However, the black curve in Figure 3.4(b) remains higher than the green one, indicating that the subject spent much more time being active on Day 1. The green curve catches up with the black one in Figure 3.4(e) at around 35% time of the day, while the corresponding green curve in Figure 3.4(b) does not. A possible explanation is that Subject 3056 performed more intense activities in late morning on Day 3 than on Day 1, with some rest in between. Since CRTA on Day 3 was reduced by rest, CRAI was equal for the two days.

3.4 Evaluation of metrics

We evaluate the validity of Activity Intensity and Time Active by comparing the metrics across subjects and activities. We then conduct an exploratory data analysis of the association between the proposed metrics with demographic factors and self report quality of life variables.

3.4.1 Validation of metrics

To validate the TA and AI metrics we focus on the replication part of the study and compare the metrics within- and between subjects for the same observed activity. In addition to the free-living data collection, the subjects were also instructed to wear the device during two lab sessions. During each session, they were asked to perform a supervised battery of activities that including: walking, stair climbing, chair-standing and lying on a bed. The start and end times of each activity were recorded by a lab technician; see [9] for a more detailed description. For a subgroup of 10 subjects we chose two types of activities, walking and chair-standing, to perform this comparison. For each lab session we chose two replicates of walking and three replicates of chair-stands. Figure 3.5 displays the raw data for these activities for Subjects 3056 and 3092. The left side of Figure 3.5 displays 4 repetitions of walking for each subject, while the right side displays 2 repetitions of chair-standing, each with 3 chair-stands.

Methods described in Section 3.2.2 were used to classify the entire time series into

"active" and "non-active". The estimated inactive periods (shaded in light gray) are highly informative as the human observer only noted when the chair-standing activity started without detailed information about exact between-activity duration. For walking, the entire period was classified as "active", as it should. We start by defining the activity intensity as "the total acceleration at a particular time point after removing global average accelerations relative to rest". Thus, any device designed to measure activity intensity in an unbiased way is a valid instrument. Of course, we do not actually have activity intensity and it is hard to check whether instruments are biased. Instead, we settle for the next best thing: checking measurement reproducibility across repetitions of the same activity and across devices.

The probability density functions of AI during walking and chair stands are shown in Figure 3.5. AI is calculated for every second and displayed as black bars under the corresponding raw-data plots. The densities of AI during walking are quite consistent within- and between-subjects. Similar results were found for all 10 subjects with in-lab data. The density curve of AI for chair-stands is different, though it displays a lot of similarity within subjects with more variability across subjects. The difference in histogram shapes between walking and chair standing is probably due to the fact that chair standing consists of three different sub-activities: resting, standing-up and sitting-down. The AI for chair-stands is low during inactive periods and high during active periods. AI during these sub-activities were quite similar within subjects across visits. In the supplementary material we also show results comparing walking

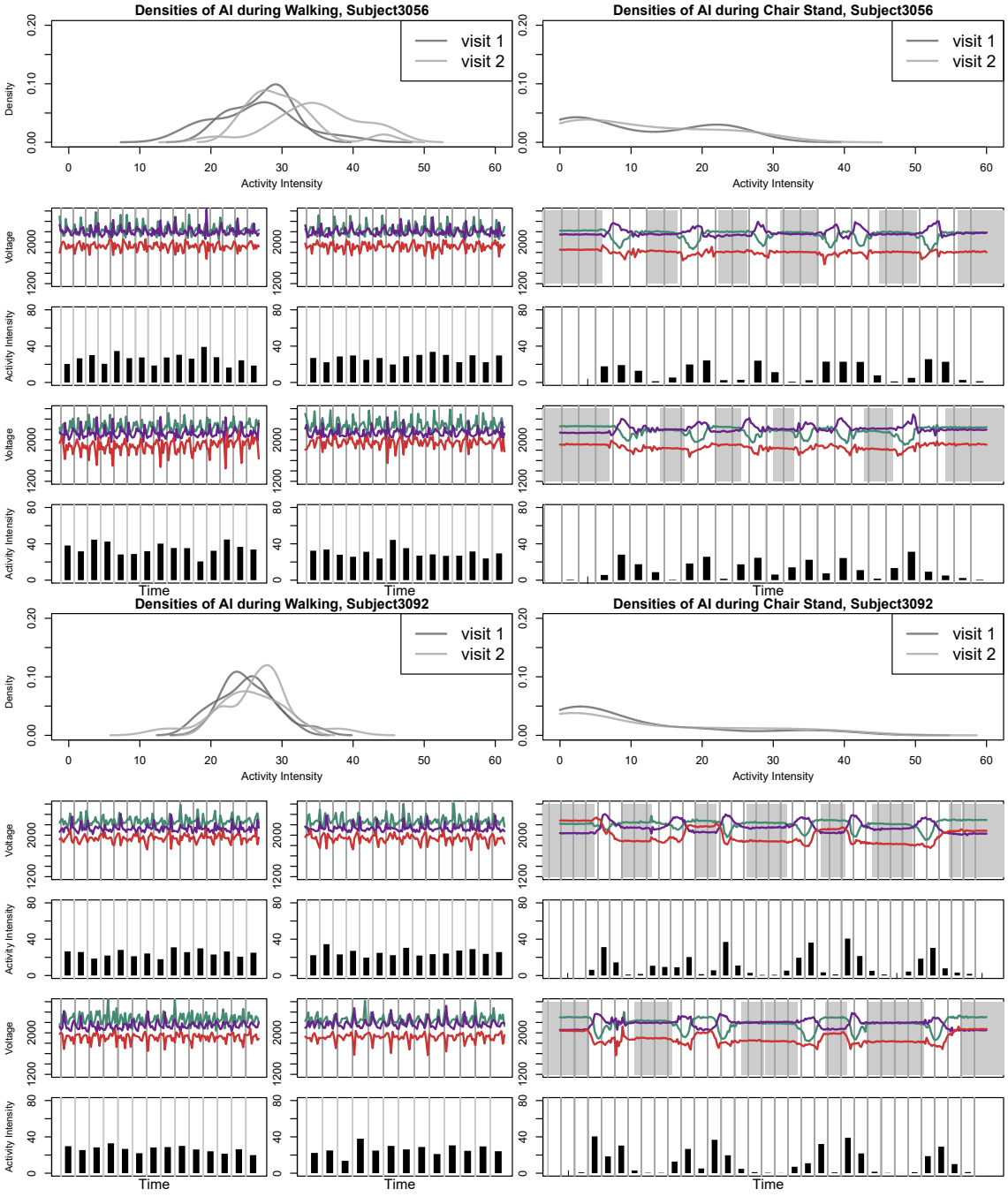


Figure 3.5: Metrics validation results for Subject 3056 and Subject 3092 during two different visits. In each visit, there are two replicates of walking and three replicates chair stands. The raw data during these periods as well as the Activity Intensity were plotted below the probability density curves of AI. In the plots of raw signals and plots of AI, each gray grid stands for one second. A lot of similarities of AI can be observed within and across subject, while the modes of the distributions are very close. Also, AI picked up the change of variability of the signal during sudden movements such as standing-up-from-chair and assigned low values to the resting periods (both standing and sitting).

normally and briskly. Results indicate that both median and standard deviation of AI increases across subjects when switching from normal to brisk walking. To quantify these differences we calculated the intraclass correlation (ICC) for walking and chair-stands. For walking the replicates are the median AI for the first and second walking period in each visit for each of the 10 subjects, respectively. The ICC for median AI for walking was 0.92. For chair-stands we manually identified the exact periods of the first, second and third replicates of chair-stands in each visit. Within each replicate we calculated the mean AI for visit 1 and 2. The ICC for mean AI for chair stands was 0.83.

3.4.2 Association with health outcomes

We now conduct an exploratory data analysis on 34 subjects from the LifeMeter study who had at least three complete days of accelerometer recordings. We investigate the possible association of Time Active mean (TAM), Activity Intensity Mean (AIM), Time Active Variability (TAV) and Activity Intensity Variability (AIV) with several different covariates: Marital Status (Marstat), Sex, Self-Reported General Health (SRH), Quality of Life (QOL), Age, Education (Edu) and Weekend. The age range of the 34 subjects (25 females: Sex= 1) was between 59 to 80, with a mean age of 68.9. Marital status is labeled as: married, separated, divorced, widowed, never married; "married" is the baseline category. Education, SRH and QOL are all treated as 0/1 variables. For education, 0 stands for having gone to high school or less (20

subjects) and 1 stands for having gone to some college or more (14 subjects). For SRH and QOL, 1 is for overall poor ratings (18 subjects for SRH and 21 for QOL) and 0 is for overall good ratings (16 subjects for SRH and 13 for QOL); "weekend" is 1 for a weekend day and zero otherwise.

Figure 3.6 displays the measurements for each of three days plotted versus covariates; results for different regression models are shown in Table 3.1 of the web supplement. Models were fit using generalized estimating equations with an exchangeable assumption for days within subject. Several significant predictors were identified: Sex (women were found to have longer time active and higher variability in intensity), Age (older individuals had lower activity intensity mean and variability), SRH (worse health status was associated with less activity) and being divorced (was associated with less activity). The weekend effect was not found to be significant (p-value > 0.5) in this data set. Separate models for women confirmed both the negative effect of worse SRH and of being divorced. We found a significant association between age and all 4 outcomes, indicating that, as age increases, both the activity level and variability decreases. Women who were never married tend to spend more time being active and exhibit a higher variation in time active. Similar results were found when SRH was replaced with QOL. The full analysis is provided in the online supplementary material.

3.5 Discussion

We provide a transparent, easy to use, and reproducible normalization approach to extract and summarize relevant metrics from raw tri-axial accelerometry data. Having a simple, explicit formula is a sine-qua-non for further refinements if the needed general discussion among researchers and users is to take place; we have provided a first step in the direction of increased transparency. Most importantly, the AI and TA measures have two built-in fail-safes: 1) using raw tri-axial accelerometer data allows future integration of data from multiple studies and platforms; and 2) using normalization with respect to sedentary and non-wear periods will likely mitigate small and moderate batch effects. Evaluating AI and other accelerometry measurements is difficult in the entire population, though validation in well defined sub-populations is probably the right approach. Our perspective is different from the current scientific practice that "acceleration is a measure of energy expenditure" or that "acceleration is a measure of a level of activity". Indeed, we consider that an accelerometer measures acceleration in three different directions at a particular part of the human body. R code is available by request and will be made available online.

3.6 Supplementary Materials

3.6.1 More on the analysis of association with health outcomes

We conduct an exploratory data analysis on 34 subjects from the LifeMeter study who had at least three complete days of accelerometer recordings. Subjects were instructed to wear the device for about 4-5 days in the free-living environment, except when they were taking a shower, swimming, or sleeping. During these times the device was taken off and placed on a table. Demographic information was collected for every subject. We investigate the possible association of Time Active mean (TAM), Activity Intensity Mean (AIM), Time Active Variability (TAV) and Activity Intensity Variability (AIV) with several different covariates: Marital Status (Marstat), Sex, Self-Reported General Health (SRH), Quality of Life (QOL), Age, Education (Edu) and Weekend. The age range of the 34 subjects (25 females: Sex= 1) was between 59 to 80, with a mean age of 68.9. Marital status is labeled as: married, separated, divorced, widowed, never married; "married" is the baseline category. Education, SRH and QOL are all treated as 0/1 variables. For education, 0 stands for having gone to high school or less (20 subjects) and 1 stands for having gone to some college or more (14 subjects). For SRH and QOL, 1 is for overall poor ratings (18 subjects for SRH and 21 for QOL) and 0 is for overall good ratings (16 subjects for SRH and

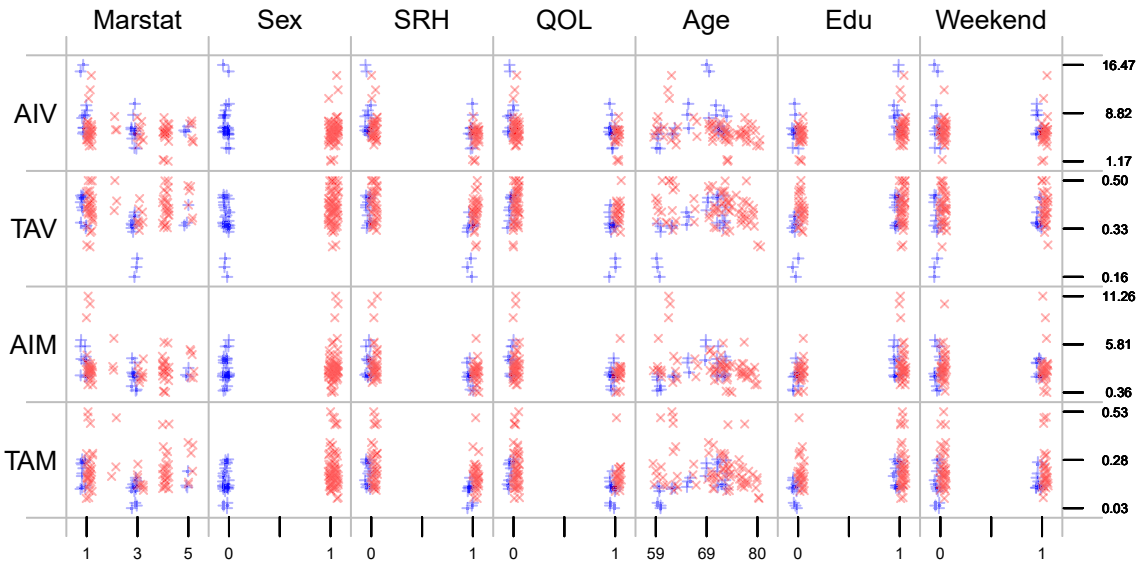


Figure 3.6: 7 different covariates (Marital Status, Sex, Self-reviewed Health, Quality of Life, Age, Education and Weekend) plotted with 4 outcomes (Time Active Mean, Activity Intensity Mean, Time Active Variability and Activity Intensity Variability). The values of covariates are slightly jittered to better reflect their relationship with the outcomes. Male and female subjects are illustrated in blue and red cross-hairs, respectively.

13 for QOL); "weekend" is 1 for a weekend day and zero otherwise.

Figure 3.6 displays the measurements for each of three days plotted versus covariates. Male and female subjects are depicted in blue and red, respectively. We started by fitting four models, each for a different outcome: TAM_{ij}, AIM_{ij}, TAV_{ij}, and AIV_{ij}. Here $i = 1, 2, \ldots, 34$ are subjects and $j = 1, 2, 3$ are days. Models were fit using generalized estimating equations with an exchangeable assumption for days within subject. Several significant predictors were identified from all or some of the models: Sex (women were found to have longer time active and higher variability in intensity), Age (older individuals had lower activity intensity mean and variability),

SRH (worse health status was associated with less activity) and being divorced (was associated with less activity). The weekend effect was not found to be significant (p-value > 0.5) in this data set.

The high correlation (0.71) between SRH and QOL may indicate over-adjustment for measures of self-reported quality of life. Thus, we refit four simpler models: one with SRH and the other with QOL. The coefficients and their p-values are shown in Table 3.1. Results indicate that: a) SRH and being divorced (Marstat= 3) are both significantly associated with all activity outcomes; b) a worse SRH is associated with lower TAM, AIM, TAV and AIV; and c) being divorced is associated with lower TAM, AIM, TAV and AIV. Age was found to be weakly associated with TAM and AIM, implying that there may be a significant decrease of average time active and activity intensity with age. We further quantify associations for women, as there were only 9 men in the sample. Results in Table 3.1 confirm both the negative effect of worse SRH and of being divorced. We found a significant association between age and all 4 outcomes, indicating that, as age increases, both the activity level and variability decreases. Women who were never married tend to spend more time being active and exhibit a higher variation in time active. Similar results were found when SRH was replaced with QOL.

Table 3.1: Estimated regression coefficients for 8 different models. Models are defined by the outcome (labeled "TAM", "TAV", "AIM", "AIV"), number of subjects (labeled as "All, $n = 34$" or "Female, $n = 25$"), and covariates (displayed under the outcome). The baseline for models labeled "All, $n = 34$" is "married subject"and for models labeled "Female, $n = 25$" is "married female".

Coefficients	Estimate	p-value		Coefficients	Estimate	p-value	
TAM: (All, $n = 34$)				AIM: (All, $n = 34$)			
Intercept	0.431	0.002	*	Intercept	8.507	0.005	*
Gender	0.036	0.022	*	Gender	0.041	0.905	
SRH	−0.068	< 0.001	*	SRH	−1.239	< 0.001	*
Age	−0.004	0.047	*	Age	−0.080	0.060	.
Separated	0.077	0.360		Separated	0.471	0.684	
Divorced	−0.096	< 0.001	*	Divorced	−1.776	< 0.001	*
Widowed	0.021	0.346		Widowed	−0.173	0.614	
Never married	0.015	0.524		Never married	−0.212	0.575	
TAV: (All, $n = 34$)				AIV: (All, $n = 34$)			
Intercept	0.436	< 0.001	*	Intercept	12.241	< 0.001	*
Gender	0.024	0.028	*	Gender	−1.546	0.015	*
SRH	−0.048	< 0.001	*	SRH	−1.637	< 0.001	*
Age	−0.001	0.293		Age	−0.073	0.107	
Separated	0.049	0.100		Separated	0.523	0.569	
Divorced	−0.064	< 0.001	*	Divorced	−1.982	0.006	*
Widowed	0.012	0.377		Widowed	−0.537	0.188	
Never married	0.008	0.597		Never married	−0.999	0.048	*
TAM: (Female, $n = 25$)				AIM: (Female, $n = 25$)			
Intercept	0.625	0.002	*	Intercept	11.308	0.009	*
SRH	−0.066	0.003	*	SRH	−1.171	< 0.001	*
Age	−0.006	0.025	*	Age	−0.118	0.038	*
Separated	0.049	0.585		Separated	−0.012	0.993	
Divorced	−0.140	< 0.001	*	Divorced	−2.463	0.005	*
Widowed	0.024	0.285		Widowed	−0.112	0.755	
Never married	0.053	0.006	*	Never married	0.489	0.069	.
TAV: (Female, $n = 25$)				AIV: (Female $n = 25$)			
Intercept	0.608	< 0.001	*	Intercept	13.821	0.001	*
SRH	−0.035	0.006	*	SRH	−1.246	< 0.001	*
Age	−0.003	0.015	*	Age	−0.115	0.044	*
Separated	0.021	0.518		Separated	−0.052	0.962	
Divorced	−0.075	< 0.001	*	Divorced	−1.991	0.036	*
Widowed	0.017	0.207		Widowed	−0.383	0.353	
Never married	0.035	0.007	*	Never married	−0.049	0.865	

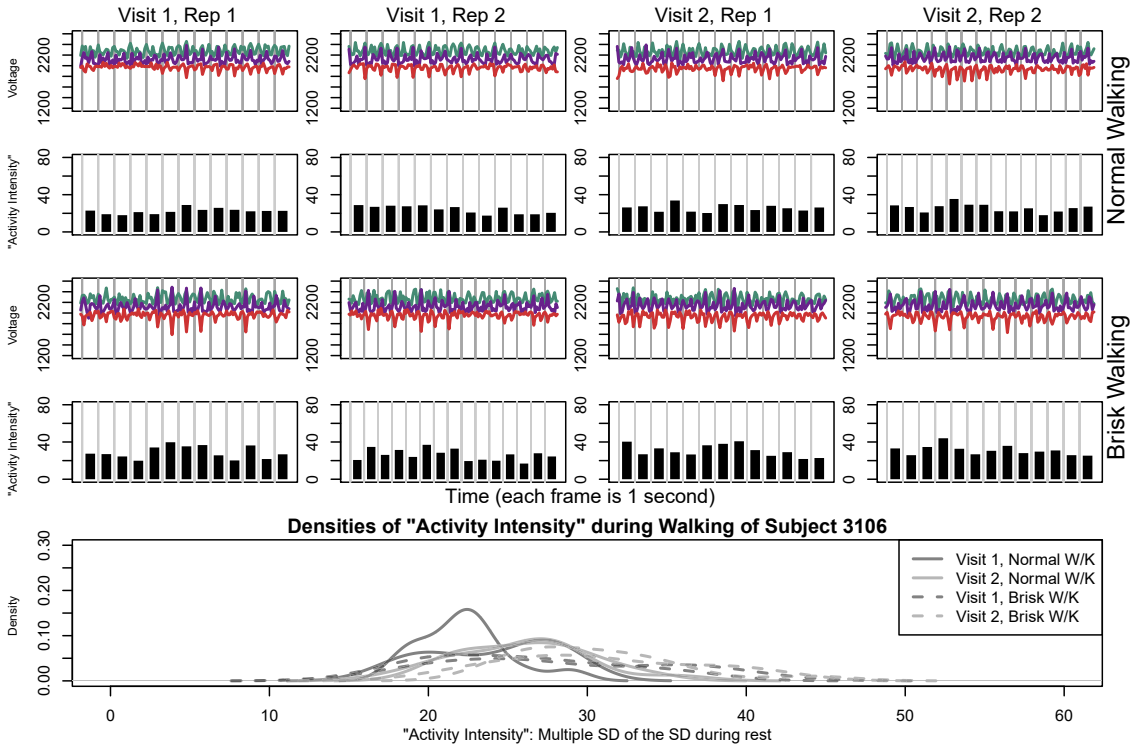


Figure 3.7: AI for subject 3106 for normal walking (raw data and AI shown in top panels) and brisk walking (raw data and AI shown in the middle panels). Histograms of AI for four repetitions of normal walking (histograms shown as solid lines) and brisk walking (histograms shown as dashed lines.)

3.6.2 More on the validation of Activity Intensity

In this section, we compare AI associated with normal walking and brisk walking in the two lab sessions discussed in Section 4.1. We chose two replicates of brisk walking from two lab sessions for 10 subjects. Figure 3.7 and Figure 3.8 display the raw data and corresponding AI for Subjects 3106 and 3208. In both figures, the first two rows display raw acceleration and AI for normal walking, whereas the following two rows correspond to brisk walking. The raw signals of brisk walking exhibit larger

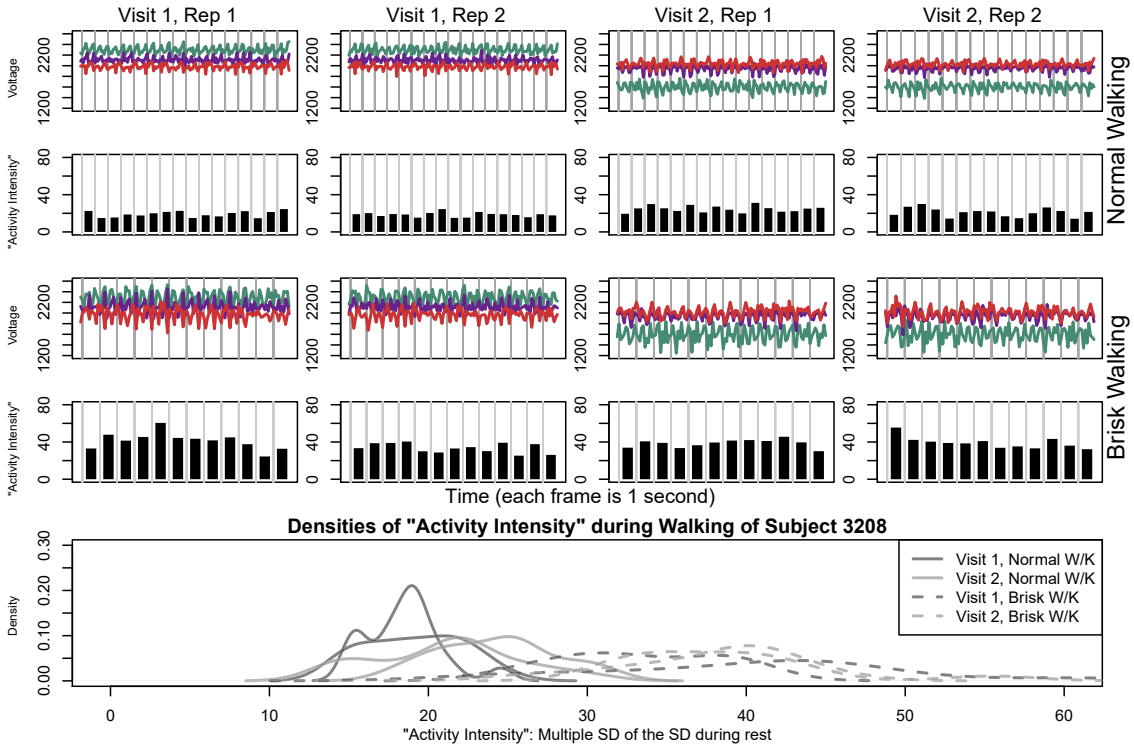


Figure 3.8: AI for subject 3208 for normal walking (raw data and AI shown in top panels) and brisk walking (raw data and AI shown in the middle panels). Histograms of AI for four repetitions of normal walking (histograms shown as solid lines) and brisk walking (histograms shown as dashed lines.)

amplitude, which leads to a larger AI, at least on average. The probability density functions for AI in Figure 3.7 and Figure 3.8 indicate these difference, with the curves associated with brisk walking being shifted to the right indicating larger average AI compared to normal walking. Interestingly, the AI distributions corresponding to brisk walking have a larger standard deviation. This could be due to a number of reasons including decreased motion control or the persistent need for low acceleration associated with re-balancing or intrinsic between-stride human walk kinematics.

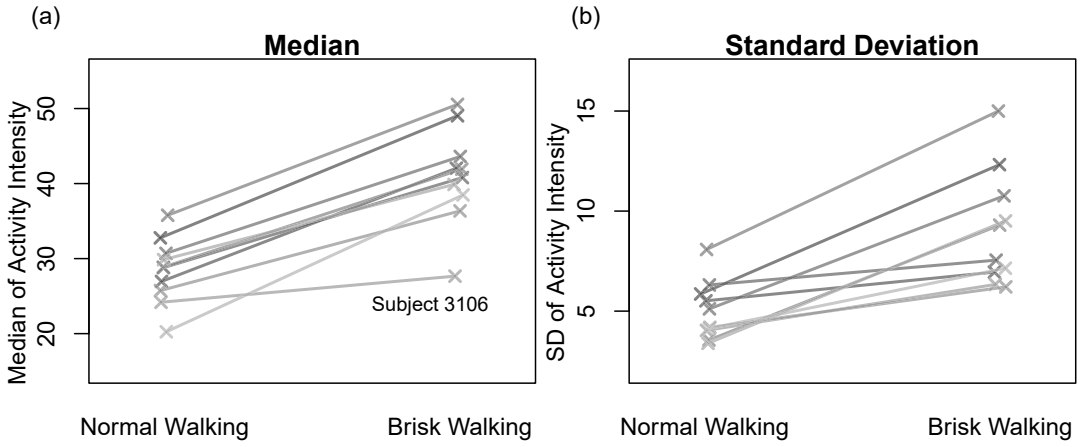


Figure 3.9: Left panel: median AI for 10 subjects during normal and brisk walking. Right panel: standard deviation of AI for for 10 subjects during normal and brisk walking.

For each subject we also computed the median and standard deviation of AI for brisk and normal walking, respectively. Figure 3.9(a) displays the median AI for the 10 subjects during normal walking (cross-hairs on the left) and brisk walking (cross-hairs on the right). The medians for the two types of walking for the same subject are connected by straight lines. Obviously, AI for brisk walking has larger median than normal walking for all 10 subjects. Figure 3.9(b) illustrates the similar comparison of standard deviation of AI. Again, brisk walking corresponds to a larger SD of AI for all subjects. However, the magnitude of the difference varies from person to person. In particular, Subject 3106 has the smallest slope in Figure 3.9(a). The corresponding density plot in Figure 3.7 indicates that separation between the density curves of AI for normal and brisk walking for this subject is not obvious. This is consistent with a careful inspection of the raw acceleration time series, which are very hard

to differentiate. This indicates that some subjects, even when told to walk briskly cannot really do so. This suggests new ways of measuring ability to walk and move.

Chapter 4

An activity index for raw accelerometry data and its comparison with other activity metrics

4.1 Introduction

Accelerometers are now commonly used to measure physical activity, and are embedded both in research and commercial devices [11, 20, 26, 48, 66, 94]. Figure 4.1 provides a conceptual analytic framework for accelerometer data in physical activity studies. While most modern accelerometers collect high-resolution signals (e.g.,

10-100 Hz), the most commonly used data output consists of summary measures over user-defined epochs (e.g., 1 minute). These measures are obtained by processing raw data using software developed by device manufacturers (see the panel "DATA TYPES" in Figure 4.1). For example, both ActiGraph GT3X+ (ActiGraph, Pensacola, FL) and Actical (Phillips Respironics, Bend, OR) software use proprietary algorithms to calculate an "activity count" (AC) [1, 48], but the two AC are not equivalent. Thus, AC has become an umbrella term for a large number of proprietary algorithms, which leads to widespread confusion among health researchers. Summary measures, such as AC, have been widely used either directly as a measure of physical activity volume or intensity, or indirectly as a predictor of energy expenditure (see analysis pathways (c), (d), (e), (f) in Figure 4.1) [12, 24, 76, 77, 85, 87].

The main reason for using such summary measures is that traditionally they were the only output of research-grade accelerometers. An important area of research is concerned with establishing the connection between summary measures of accelerometry and standard measures of physical activity intensity (metabolic equivalents (MET)) and physical activity volume (MET-min) (see review [91]). One goal of that research is to translate accelerometry summaries into physical activity intensity categories: sedentary, light, moderate, and vigorous activity. Thus, there are many calibration studies designed to identify the cut-points for physical activity counts that correspond to intensity categories [31, 35, 36, 72, 75, 85], as well as studies designed to translate count data into METs and caloric expenditure [20, 25, 81].

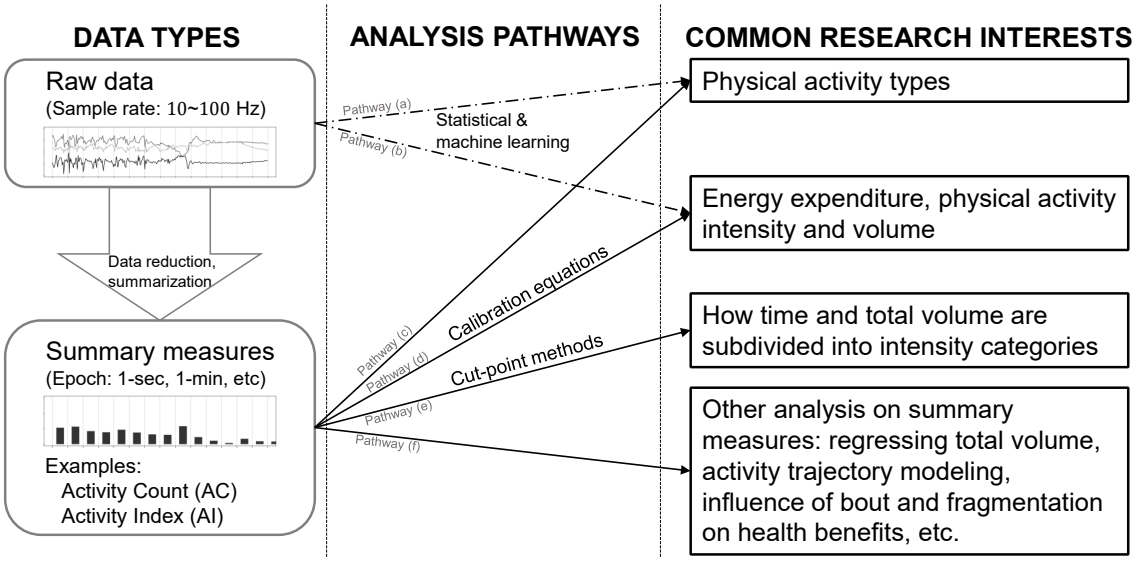


Figure 4.1: A general framework for accelerometer-related studies. The left panel illustrates two general data types: raw data and summary measures. The right panel shows 4 common research interests. The mid panel contains 6 common analysis pathways between the data and the research interests.

More recently, high-resolution raw accelerometry data has become available on various devices, including on the ActiGraph GT3X+ and GENEActiv (Activinsights Ltd, Cambridgeshire, UK) accelerometers. Rather than relying on manufacturer software, researchers have started to develop new analytic approaches for the raw data. Of particular interest has been physical activity type recognition (pathway (a), [9, 13, 28, 42, 73, 80]) and energy expenditure or MET prediction (pathway (b), [28, 80]) using statistical and machine learning approaches. However, little research has been focused on developing an explicit, open-source, and reproducible summary metric based on raw data as an alternative to existing metrics (e.g., AC). The need for such measures is central to physical activity research, as current definitions of AC

are proprietary and device- and software-specific [1, 10]. A transparent and publicly available summary metric derived from raw data has the potential to allow comparisons of results across studies that use different accelerometers, improve translation among studies, and allow a more uniform interpretation of results.

Two notable summary metrics based on raw accelerometry data are Activity Intensity (AI_0) by Bai et al. [10] and Euclidean Norm Minus One (ENMO) by van Hees et al. [90]. The AI_0 measures the amplitude of the raw accelerometry signal relative to its amplitude distribution at rest, while ENMO is the vector magnitude of raw signals after removing $1g$ (one Earth standard gravitational unit). Both AI_0 and ENMO are designed to quantify the magnitude of acceleration during a given epoch. AI_0 has a publicly available formula and clear interpretation. However, its reliance on the choice of inactive periods and a threshold for systematic noise make it relatively difficult to implement in large studies. ENMO was also reported to be highly associated with physical activity energy expenditure [44, 90], but it was not directly compared with AC. In this paper, we propose a new physical activity index (AI), which substantially improves AI_0 by reducing its reliance on identifying all rest periods, making it rotationally invariant, and ensuring the consistency of definition across time domains. We show that AI outperforms both AC and ENMO in terms of prediction of physical activity energy expenditure and classification of physical activity intensity.

4.2 Materials and Methods

4.2.1 Participants

The Objective Physical Activity and Cardiovascular Health (OPACH) Study is an ancillary study of the Womens Health Initiative 2010-2015 Long Life Study. The OPACH included a calibration sub-study, where 200 women aged 60 to 91 years old were invited to participate in one laboratory session to calibrate accelerometry counts to energy expenditure. This sub-study was approved by the Institutional Review Boards from each data collection site and by the Womens Health Initiative Clinical Coordinating Center. Participants were asked to visit the study clinic site where they signed an informed consent form and completed a brief questionnaire.

4.2.2 Accelerometry

ActiGraph GT3X+, a tri-axial accelerometer, was used in OPACH to measure physical activity. It was set to collect 30 Hz raw acceleration time series (x, y and z axes). GT3X+ features an "idle sleep mode" [2], which means when an internal algorithm detects no movement for 10 consecutive seconds, the last sampled raw acceleration value during the 10th second is repeated infinitely until movement is detected by the algorithm again. The ActiLife 6 companion software could then calculate axis-specific AC (AC_x, AC_y and AC_z for 3 axes) and the AC Vector Mag-

nitude ($\sqrt{\mathrm{AC}_x^2 + \mathrm{AC}_y^2 + \mathrm{AC}_z^2}$), which is the square root of the sum of square of the axis-specific AC) using the raw acceleration time series. The AC Vector Magnitudes are referred to as AC in the rest of this manuscript. An optional data processing procedure called "Low Frequency Extension (LFE)" [3] was also implemented while calculating AC. This is a variation on the AC measurement designed to improve ACs sensitivity to sedentary and light activities [21]. AC with and without LFE were both calculated and used in this paper.

4.2.3 Data collection

The participants performed several standardized tasks while simultaneously wearing an accelerometer, a heart rate monitor, and a portable indirect calorimeter to measure oxygen uptake. The hip-worn accelerometer was placed at the iliac crest and secured with a belt. Oxygen uptake (VO_2) and heart rate were measured continuously during the physical activity tasks using the Oxycon Mobile (CareFusion, Rolle, Switzerland), a portable, battery operated, breath-by-breath metabolic unit.

The tasks of the calibration study were selected to vary in intensity from sedentary to moderate intensity in older women. Women provided Borg ratings of perceived exertion (RPE) [15] for each task to ensure level of effort did not exceed moderate-intensity. With the exception of treadmill walking at different speeds, participants rested $\geq$ 2 minutes between activities so that heart rate could return to within 10 beats/minute of resting heart rate. Simultaneous measurements of accelerometer

counts, heart rate, and VO_2 were recorded during the entire period for each physical activity. The duration of tasks was chosen to achieve steady rate metabolism for measurement of task-specific oxygen uptake. The participants performed tasks in the following order: watching DVD while sitting quietly (alias: DVD), assembling puzzle while sitting (alias: PUZZ), washing dishes while standing (alias: DISH), doing laundry while standing (alias: LAUD), 400-meter walking (alias: WALK), dust mopping while standing (alias: MOP), treadmill walking at 1.5mph (alias: TM15), and treadmill walking at higher speed, either 2.0mph (alias: TM20) or 2.5mph (alias: TM25) depending on the RPE. Determination of a 2.0 mph vs. a 2.5 mph pace for the second walking stage was based on participants RPE after 5 minutes into the 1.5 mph walk. At this point, women reporting a RPE of $\leq$ 11 walked at the 2.5 mph pace, while those reporting a RPE of 12-14 walked at the 2.0 mph pace. Women with a RPE $\leq$ 14 did not continue with the faster paced treadmill walk.

In the rest of the paper, the physical activity types are referred to by their aliases. The raw accelerometry data were used to compute AI and ENMO. The VO_2 was converted to average energy expenditure during each activity in METs, by dividing the oxygen intake by 3.5 mL/(kg·min). In addition, standard measurements such as weight, height and blood pressure were taken during the laboratory visit. More details about these measurements and protocol can be found elsewhere [31].

4.2.4 The new Activity Index

Raw accelerometer data measure total acceleration from both device movement and gravity; the latter of which is always $1g$ downward. As previously reported [10], the variability of raw acceleration signals (standard deviation or variance) in short epochs (e.g., 1 second) removes gravity and provides a summary measure of movement intensity. The standard deviation captures the magnitude of the signals oscillation. When the frequency of such oscillation increases (e.g., when the accelerometer wearer switched from walking to running), the standard deviation can detect the increased variability of the signals, while the mean may not change accordingly. Thus, we chose to use the variance of raw accelerometry data along the three axes as building blocks to construct the proposed metric.

Specifically, let $\sigma^2_{im}(t; H)$ denote the variance of participant is acceleration signals along axis m ($m = 1, 2, 3$) in the window of length H starting at t. We then aggregate the variability of three axes by taking their sum, $\sigma^2_{i1}(t; H) + \sigma^2_{i2}(t; H) + \sigma^2_{i3}(t; H)$. The sum was normalized using the systematic noise variance denoted by $\bar{\sigma}^2_i = \sigma^2_{i1} + \sigma^2_{i2} + \sigma^2_{i3}$, so that it yields zero values when the device is not moving. $\bar{\sigma}^2_i$ depends on the accuracy of the device and can be calculated using raw data in periods while the accelerometer is not moving. Specifically, σ^2_{im} ($m = 1, 2, 3$) is the average of $\sigma^2_{im}(t; H); t \in \mathcal{T}_i$ where $\mathcal{T}_i$ stands for collection of time points t when the accelerometer is not moving. The proposed Activity Index, $\mathrm{AI}^{\mathrm{ABS}}_i(t; H)$, for an epoch of length H starting at time t is

defined by

$$\mathrm{AI}_i^{\mathrm{ABS}}(t;H) = \sqrt{(\max\left[\frac{1}{3}\sum_{m=1}^{3}\left\{\sigma_{im}^2(t;H) - \bar{\sigma}_i^2\right\}, 0\right]}. \tag{4.1}$$

The AI captures the variability of device acceleration in excess of systematic noise and has the same unit as "g".

In practice, we found that the AI is typically in a narrow range, as expressed in unit of "g", especially for sedentary to light activities. To enhance interpretability, we also present a modified version of AI on a relative scale. Specifically, we further standardize the AI using the systematic variance $\bar{\sigma}_i^2$,

$$\mathrm{AI}_i^{\mathrm{REL}}(t;H) = \sqrt{(\max\left[\frac{1}{3}\left\{\sum_{m=1}^{3}\frac{\sigma_{im}^2(t;H) - \bar{\sigma}_i^2}{\bar{\sigma}_i^2}\right\}, 0\right]}, \tag{4.2}$$

so that an AI value of 1 is equivalent to the smallest amount of variability detectable by the device. The values of the relative scale AI spread in a wider range similar to AC, and might be preferred by some researchers. For studies that utilize the same accelerometry device for all participants, two versions of the AIs are directly proportional, with the constant of proportionality equal to $\bar{\sigma}_i^2$, so their performances are equivalent. In the application to the OPACH study, we reported results using the relative scale AI in the application for ease of presentation and interpretation. Detailed definitions of all these quantities are provided in the Supplementary Materials in Section 4.5.

The newly proposed AI has three desirable properties: ease of implementation, additivity, and rotational invariance. As these properties hold for AI in both absolute and relative scale, we denote the new AI as $\mathrm{AI}_i^{\mathrm{new}}(t;H)$ regardless of its

scale. With an explicit formula, $\text{AI}_i^{\text{new}}(t; H)$ could be implemented in a computationally efficient way for large epidemiology studies with tens of thousands of participants wearing accelerometers. For additivity, we defined the second-by-second $\text{AI}_i^{\text{new}}(t; H_s)$ to be the finest level for computing AI, where H_s was the window size for one second. Any aggregated AI (e.g., 1-minute AI or $\text{AI}_i^{\text{new}}(t; 60H_s)$) was obtained by summing up all the adjacent 1-second AIs within that period (e.g., $\text{AI}_i^{\text{new}}(t; 60H_s) = \sum_{s=1}^{60} \text{AI}_i^{\text{new}}(t + s - 1; H_s)$). Rotational invariance means that the AI summarizes the magnitude of movement over three axes, regardless of whether the orientation of the device.

These properties are described in the discussion, while the technical details and proofs are included in the Supplementary Materials in Section 4.5. Note that although previously proposed AI_0 [10] was also based on standard deviation of acceleration signals, it had several drawbacks and did not possess the three properties discussed above. Specifically, AI_0 requires a participant-specific tuning parameter for the metric normalization, while AI only requires a device- or study-specific tuning parameter. Unlike AI, AI_0 does not guarantee rotational invariance because it combines variability from 3 axes using the sum of standard deviations instead of the sum of variances. More details on the difference between AI and AI_0 is included in the Supporting Information.

4.2.5 Statistical analysis

4.2.5.1 Data processing

Among 200 women from the OPACH calibration study, 194 had complete raw accelerometry data available, which were used in our analysis. Second-by-second AI was computed for each participant during each physical activity. Due to the "idle sleep mode" of the ActiGraph, the 10-second periods in the beginning of the selected non-wear (idle) periods were used to estimate $\bar{\sigma}_i$. We calculated $\bar{\sigma}_i$ for 10 participants and found them to be very close to each other. Such consistency of $\bar{\sigma}_i$ across different participants (or essentially, devices) allowed us to combine them into $\bar{\sigma} = \sum_{i=1}^{I} \bar{\sigma}_i$, which was a study-specific systematic variation. Using $\bar{\sigma}$, second-by-second $\mathrm{AI}_i^{\mathrm{new}}(t; H_s)$ was computed for Participant i at time t. Second-by-second ENMO was computed by calculating the average of $\left\{\max\left[0, \sqrt{\sum_{m=1}^{3}\{X_m(t+s-1)\}^2}\right] s = 1, 2, \ldots, H_s\right\}$ during each one-second window $[t, t + H_s - 1]$ [44], where $X_1(t)$, $X_2(t)$ and $X_3(t)$ are the raw acceleration signals of each axis and H_s is the window size for one second. The corresponding AC and AC with LFE at each second were computed using the ActiLife software.

4.2.5.2 Directly comparing AI, AC and ENMO

Second-by-second AI, AC, and ENMO measurements were compared using different approaches. First, scatterplots of AI versus AC, and AI versus ENMO, for

randomly selected participants were explored. Second, for each of AI, AC, AC (LFE), and ENMO, a boxplot of pooled metrics across all participants was generated for all 9 physical activities. Third, receiver operating characteristic (ROC) analyses were conducted to assess and compare the performance of AI, AC and ENMO in distinguishing different activity types. More specifically, we illustrated comparisons with examples of 4 pairs of activities: DVD vs. DISH, DVD vs. LAUD, DVD vs. PUZZ and WALK vs. MOP. The area under the ROC curves (AUC) was used to evaluate the prediction performance of each measurement, as it represents the accuracy of the test to discriminate between two samples, with values significantly greater than 0.5 indicating better discrimination than by chance alone.

4.2.5.3 Comparing MET prediction performance of AI, AC and ENMO

We compared AI, AC, and ENMO in terms of their predictive performance of energy expenditure, as measured by a portable indirect calorimeter in METs. Median METs during each activity were analyzed together with median AI, AC, and ENMO during each activity type. Scatterplots of AI, AC, and ENMO versus METs were used for visual inspection of these associations, with Pearson correlation coefficients reported. We also evaluated the performance of AI, AC, and ENMO when differentiating between activities of different intensities as defined by thresholds on METs. Sedentary behaviors were defined as those with MET$<$ 1.5, light activities as

those with MET$\in [1.5, 3)$ and moderate-to-vigorous activities as those with MET3. AUC was used to compare the prediction performance of these metrics to distinguish between activities performed at different levels of energy expenditure.

Software. AC and AC (LFE) were computed using ActiLife (version 6.11.8; ActiGraph, Pensacola, FL). AI and ENMO computation, as well as the statistical analysis were performed in R (version 2.15.3; R Foundation for Statistical Computing, Vienna, Austria). The R package for AI computation is available on GitHub `https://github.com/javybai/ActivityIndex`.

4.3 Results

4.3.1 Summary statistics

The 194 women used in our analysis had a mean age of 75.4 years (standard deviation 7.7), with 21.6% ($n = 42$) between 6069 years, 44.8% ($n = 87$) between 7079 years, 31.4% ($n = 61$) between 8089 years, and 2.1% ($n = 4$) between 9091 years. For body mass index, the participants were evenly distributed across normal weight, overweight, and obesity categories ($n = 68$, 60 and 63, respectively), while 3 participants were underweight.

Summary statistics (mean and standard deviation) for AI, AC, AC (LFE), and ENMO from the study are shown in Table 4.1. For sedentary, light, and moderate activities, the mean of AI and both ACs increased in the order of the energy cost

Table 4.1: Summary statistics of second-by-second AI, AC, AC (LFE), and ENMO of each activity.

Activity Type	**Subject Size**	**Mean (SD) AI**	**Mean (SD) AC**	**Mean (SD) AC(LFE)**	**Mean (SD) ENMO**
DVD	194	0.61(3.03)	0.42(5.06)	0.53(5.35)	0.001(0.004)
PUZZ	193	5.20(5.69)	2.21(9.20)	3.39(10.04)	0.001(0.004)
DISH	194	9.33(7.66)	4.53(12.29)	6.43(13.17)	0.002(0.007)
LAUN	194	13.53(9.73)	12.15(19.79)	15.38(20.19)	0.002(0.006)
MOP	193	25.57(13.55)	29.69(25.22)	33.73(24.66)	0.011(0.015)
WALK	190	61.88(19.78)	43.49(19.41)	48.06(18.94)	0.073(0.033)
TM15	171	41.62(10.37)	24.61(16.15)	29.80(15.41)	0.041(0.018)
TM20	53	49.30(13.30)	31.02(19.45)	36.18(18.75)	0.052(0.021)
TM25	90	63.60(15.08)	42.02(15.59)	46.62(15.19)	0.076(0.025)

of the activities: DVD, PUZZ, DISH, LAUN and MOP. However, the mean ENMO for the first four activities were similar, suggesting that ENMO may underperform the other metrics in terms of distinguishing between types of sedentary and light activities. For the three treadmill walking speeds, all metrics performed as expected, increasing as the speed increased from 1.5mph to 2.5mph. However, the ratio of the mean divided by the standard deviations of ENMO and both ACs were substantially larger than that of AI for each activity, indicating smaller heterogeneity for AI.

4.3.2 Directly comparing AI, AC and ENMO

Figure 4.2 displays second-by-second scatterplots of AI (y-axis) versus AC (x-axis) (Figure 4.2A) and AI (y-axis) versus ENMO (x-axis) (Figure 4.2B) for a randomly selected participant. The dots were rendered in different colors to distinguish among different activity types (one color per activity). To reduce over-plotting, we randomly sampled 100 seconds from each activity and only displayed the AI, AC and ENMO during these sampled seconds. The figure shows that ACs and ENMOs were often equal or very close to 0 for sedentary behaviors such as DVD and PUZZ. For light intensity activities including DISH and LAUN, AC displayed a wide spread in the range 060 with many zero values, while AI values were mostly nonzero and tended to be more clustered for each activity type. ENMO was highly correlated with AI for moderate activities (MOP, WALK, TM15, TM20 and TM25).

Figure 4.3 illustrates the distribution of AI, AC, AC (LFE), and ENMO (after pooling all participants) for each activity. It confirmed that both AC and ENMO (Figure 4.3B, 4.3D) had values very close to 0 for sedentary and light activities (such as DVD, PUZZ, DISH, and LAUN). Though LFE increased ACs sensitivity to sedentary and light activities (Figure 4.3C), there were still substantial zero counts for DVD, PUZZ, and DISH (with median close to zero). In contrast, AI in Figure 4.3A displayed distributions that were more separable for different activities, as the median AI values increased with activity intensity. For high light to moderate intensity physical activities, such as 400-meter walking and treadmill walking, the values of all

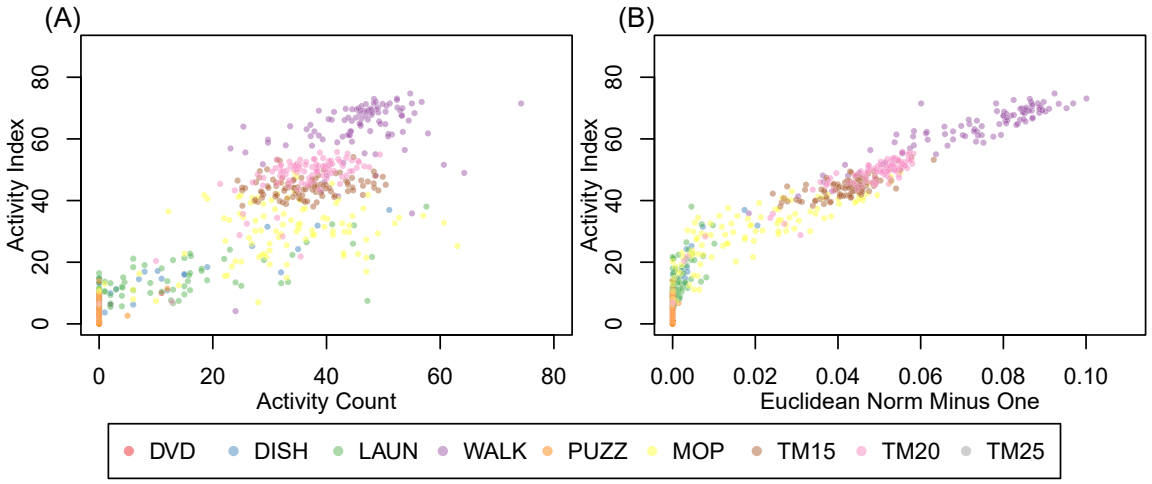


Figure 4.2: Scatterplots of Activity Index (AI, y-axis) versus activity count (AC, x-axis) (A) and AI (y-axis) versus Euclidean Norm Minus One (ENMO, x-axis) (B) for a randomly selected participant. Each point representing activity summery metrics in a 1-second interval. The points were rendered in different colors to represent different activity types. A random sample of 100 seconds were shown for each activity to reduce over-plotting.

four metrics were more concentrated and increased with gait speed. These observations implied that AI provides summary metrics for raw accelerometry signals that are more likely to be distinguishable among activities.

Figure 4.4 displays the four ROC plots of distinguishing various types of sedentary to light activities using AI, AC and ENMO. The solid, purple dashed, orange dashed, and dotted curves were ROC curves of AI, AC, AC (LFE), and ENMO, respectively. The dashed and dotted curves in Figure 4.4A and 4.4C were closer to the diagonal line (equivalent to a random guess classifier), while the corresponding solid curve was much higher overall. It indicated that neither AC nor ENMO could effectively differentiate DVD from DISH or PUZZ (AUC smaller than/close to 0.50), whereas AI

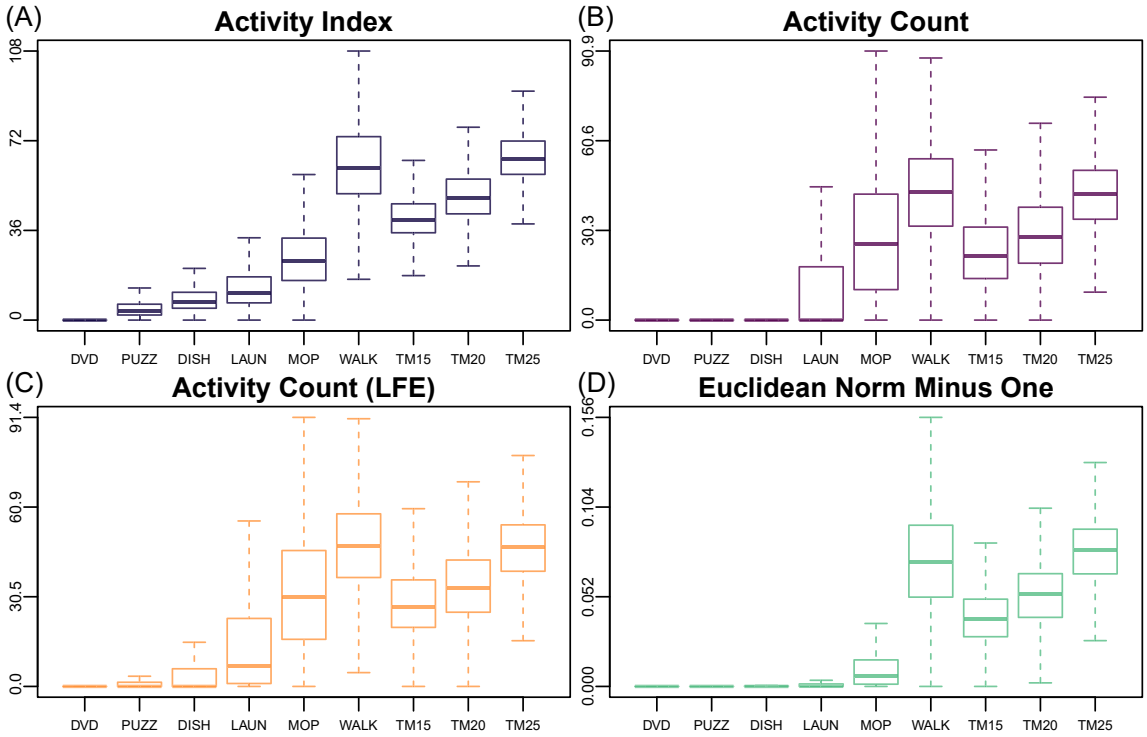


Figure 4.3: Comparison of the boxplots of Activity Index (AI), activity count (AC), AC with Low Frequency Extention (LFE) and Euclidean Norm Minus One (ENMO) during different types of activities. Outliers outside of the upper and lower whiskers are omitted. Each type of summary metric from all the participants were pooled together and plotted according to the type of activity.

performed much better with an AUC greater than 0.90. The predictive performance of AC and ENMO increased with METs (in the order of PUZZ, DISH and LAUD), but AI had substantially higher AUCs in all cases. In general, AC with LFE had greater AUC than both AC without LFE and ENMO, corresponding to better predictive performance for sedentary and light activities (Figure 4.4A, 4.4B and 4.4C). Figure 4.4D displays the performance of AI and AC for a pair of moderate to vigorous physical activities (MVPA), MOP versus WALK. For this pair of activities of the pre-

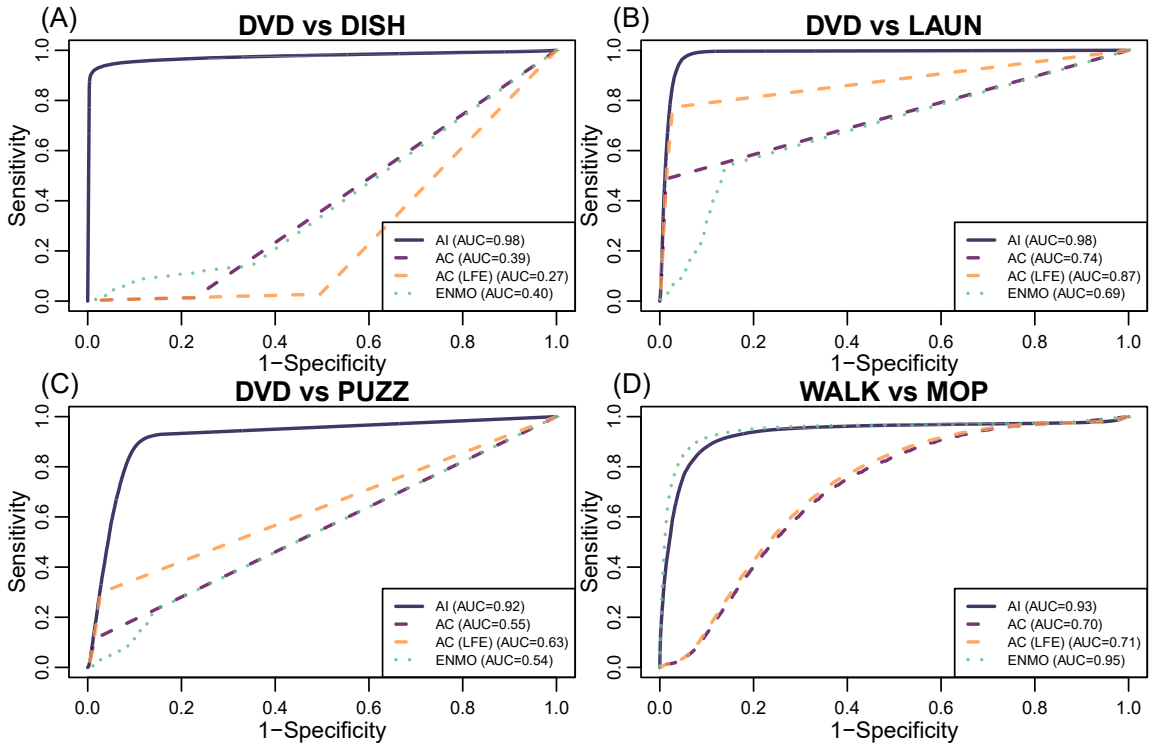


Figure 4.4: The "receiver operating characteristic" (ROC) curves for distinguishing four pairs of activity types, using Activity Index (AI, solid curves) or activity count (AC, dashed curves in different color for AC with and without Low Frequency Extension (LFE)) or Euclidean Norm Minus One (ENMO, dotted curves), respectively. The corresponding area under the curve (AUC) of each ROC curve is given in the legend section.

diction performance of AC with and without LFE was very close (AUCs $\sim$ 0.700.71). This suggests that LFE enhanced the prediction performance of AC for sedentary to light activities but not for moderate activities. AI and ENMO out-performed both versions of AC in this case, with an AUC of 0.93 and 0.95, respectively. Figure 4.4 confirmed that AI provided most distinguishable summary metrics for activities at every level of activity intensity, while AC and ENMO performed well only for MVPA. In addition, ENMO performed as well as AI for MVPA.

4.3.3 Comparing MET prediction performance of AI, AC and ENMO

Figure 4.5 shows scatterplots of METs versus four metrics: AI, AC, AC (LFE) and ENMO. The METs were positively correlated with all four metrics, with coefficients of determination (R^2) values of 0.72, 0.54, 0.59 and 0.62 for AI, AC, AC (LFE) and ENMO, respectively. Although the ACs and ENMOs were correlated with METs, they were close to 0 for DVD, DISH and PUZZ, while ENMO was close to 0 even for LAUN. In contrast, the MET values for these activities were different, suggesting that ENMO and AC may underperform in terms of predicting low intensity activities. The AC (LFE) exhibited slightly improved sensitivity to sedentary and light activities. In contrast, AI tracked the increase in METs much closer for all activities.

ROC analyses were conducted to further quantify these findings. Figure 4.6 provides the ROC curves of AI, AC (with and without LFE) and ENMO to classify activity intensity categories such as sedentary (< 1.5 METs), light (1.5-3 METs) and MVPA (> 3 METs). AI performed better than AC for all activities, while ENMO had slightly worse performance than AC for sedentary and light activities (Figure 4.6B and 4.6C, with AUC 0.85 v.s. 0.86 and 0.74 v.s. 0.75). ENMO performed very well when differentiating MVPA and other activities (Figure 4.6C), with a AUC comparable to that of AI (0.97 v.s. 0.96). The predictive performance of both versions of AC and ENMO was better for MVPA versus light activities (Figure 4.6A) than light

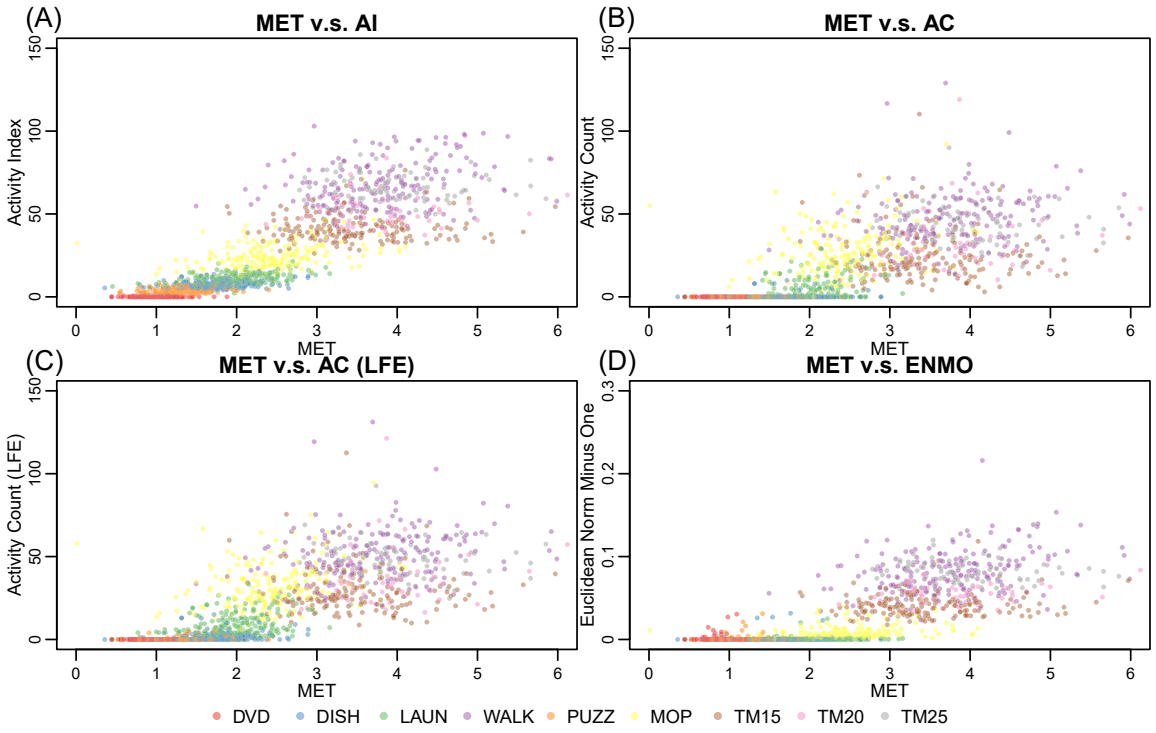


Figure 4.5: Scatterplots of metabolic equivalents (METs) versus Activity Index (AI) (A), activity count (AC) (B), AC with Low Frequency Extension (LFE) (C) and Euclidean Norm Minus One (ENMO) (D). MET is on x-axis for all four plots, while AI, AC, AC (LFE) and ENMO are on the y-axis in (A), (B), (C) and (D), respectively. Each point in the figure represents a participant's median METs during a certain activity (rendered in different colors) versus the median AI, AC or ENMO while he/she was performing the same activity.

versus sedentary activities (Figure 4.6C). This indicates that both AC and ENMO are severely limited as classifiers of sedentary and light activities. The AC (LFE) performed better than AC for distinguishing between sedentary and light activities (AUC increased from 0.75 to 0.85 in Figure 4.6C). The AUC for predicting light versus MVPA was about the same for AC with and without LFE (both 0.92 in Figure 4.6A). This indicated that LFE does not substantially improve the predictive performance

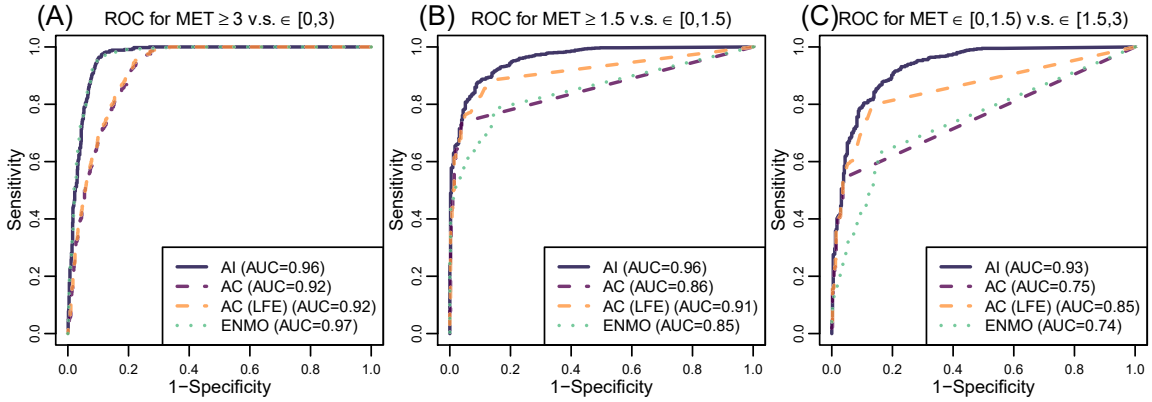


Figure 4.6: The "receiver operating characteristic" (ROC) curves of Activity Index (AI), activity count (AC), AC with Low Frequency Extension (LFE) and Euclidean Norm Minus One (ENMO) to predict whether metabolic equivalents (MET) is smaller or greater than 3 (A) and 1.5 (B), and whether MET is bigger than 1.5 but smaller than 3 (C). The ROC curves for AI, AC and ENMO are solid, dashed and dotted, respectively, while AC with and without LFE are rendered in purple and orange. The corresponding area under the curve (AUC) of each ROC curve is given in the legend section.

of AC for MVPA.

4.4 Discussion

We proposed AI, a new metric of physical activity based on high-resolution raw accelerometer data. The AI has several desirable properties including transparency, ease of deployment, additivity, and rotational invariance. The new metric was compared to the established AC (with and without LFE) and ENMO using laboratory data from 194 women 60-91 years of age in the OPACH Study. We found that the AI was the best in distinguishing among various types of physical activities across different intensity levels. AI had the best overall performance in terms of predicting

energy expenditure expressed in METs, and had better predictive performance for classifying an epoch into various physical activity intensity categories.

As the systematic noise $\bar{\sigma}_i$ for AI computation is determined using a pre-annotated "non-wear period", AI could be implemented in large epidemiology studies with low effort. The non-wear periods may be obtained in various ways. First, participants could report non-wear periods. Second, data could be collected while accelerometers are placed on a desk, before being used in the study. In this case, the standard deviations, $\bar{\sigma}_i$, are accelerometer-specific, instead of participant-specific. Therefore, only one $\bar{\sigma}_i$ needs to be computed for each accelerometer used in the study, and the same $\bar{\sigma}_i$ can be used for all participants who use the same accelerometer. Third, a published algorithm [22, 23] could be used to identify non-wear periods. Moreover, as in our data analysis, we could further combine all the standard deviations, $\bar{\sigma}_i$, into a "study-specific" standard deviation, $\bar{\sigma}$, and use this single parameter throughout a study. As long as $\bar{\sigma}_i$'s are not very different across accelerometers, this approach is reasonable. A simple histogram of $\bar{\sigma}_i$ can indicate whether the assumption is valid in the study and could identify miss-calibrated accelerometers.

The additivity and rotational invariance are both desired measurement properties that define the new AI as a proper physical activity measure. AI is additive in the sense that AI values in different epochs can be added to provide an aggregated AI that is consistent across resolutions. Additivity is an important self-consistency feature, as it ensures that AI is comparable and generalizable across studies. For

instance, if Study A suggests that people use 30 AI per second as the cut-off for light intensity and moderate intensity activities, it is equivalent to suggest $30 \times 60 = 1800$ AI per minute as the cut-off in Study B that calculates AI per minute. Moreover, AIs rotational invariance guarantees it remains unchanged while the participant is performing the same type of physical activity with a rotated accelerometer placed at the same location on the body. This property is crucial in practice. In many studies that collect free-living data the device can rotate or tilt when they are being equipped or during the data collection. Our proposed AI theoretically guarantees rotational invariance and reduces noise and bias due to rotation in practical applications. We expect that this will translate in better robustness to rotations and small changes in location on the body.

Although summary measures like AI, AC, and ENMO do not retain all the information in the raw data, analyses based on these metrics should remain a major part of research, due to the substantially reduced data size and explicit interpretation. Figure 4.1 indicates that the majority of analysis pathways rely on summary measures. Our proposed AI is an open-sources alternative to the popular AC for summarizing the raw data, which is a crucial bridge between the raw data and summary measures. To demonstrate AI is indeed a better method than others to summarize raw data, we showed the AI yielded to more distinct values than AC and ENMO based on the raw data of different activities. AI was also more highly correlated with METs than AC and ENMO, and performed much better while used to classify activities of different

intensity categories. Both Evenson et al. [31] and our study showed that although ActiGraph attempted to improve AC using LFE to better capture low-amplitude movement [33], the improvements over standard AC are modest. ENMO is another important open-source summary metric for raw accelerometry data, but was proved to be outperformed by AI for sedentary and light activities. We contend that replacing or complementing AC with AI would provide much-needed transparency for raw data processing and will greatly enhance the characterization of sedentary and light activity.

While the comparison of AI with other metrics was conducted using data from a group of older women, AIs advantage over AC and ENMO is not limited to this population. Indeed, the approach summarizes information contained in the acceleration time series data, which are independent of population characteristics. While we have demonstrated that AI outperforms AC and ENMO in terms of quantifying sedentary and light activities, it also performs better (than AC) or equally well (as ENMO) for MVPA. Therefore, while populations of other ages also perform activities in these four intensity categories and produce similar acceleration time series, we expect that, in those cases, AI will still have better performance at least for sedentary and light activities. Future studies on youth and young adults, employing a range of physical activity intensities, can explicitly test this.

Our work has several potential limitations. For example, only one type of accelerometer, the ActiGraph GT3X+, was considered in this study. It remains an

open problem to compare the AI collected from other devices, though we expect consistent results for well-calibrated accelerometers. Shaker studies or studies using several devices simultaneously could be conducted to answer this type of question. Another limitation is that we only investigated data from hip-worn accelerometers. As many current studies have moved towards wrist-worn accelerometers to improve compliance, it is important to understand how AI performs for wrist-worn accelerometers. A final noteworthy limitation is our focus on women age 60 years and older. Exploration in other samples is warranted. Nevertheless, the proposed AI provides a novel and transparent way to summarize densely sampled raw accelerometry data, and may serve as an alternative to AC.

4.5 Supplementary Materials

We first illustrate in detail how Activity Index (AI) is formalized. Then, we discuss the three properties of AI more rigorously.

4.5.1 Definitions and mathematical formula of AI

We first introduce notations. Denote the data by $\mathbf{X}_i(t) = \{X_{i1}(t), X_{i2}(t), X_{i3}(t)\}$ $(t = 1, 2, ..., T_i)$, where T_i is the length of the accelerometer time series for Participant i. Let f denote the sample rate ($f = 30$Hz in our study) and H be the window size of raw data to be summarized into one AI measure, where $H \geq f$ (windows size is no less

than the sample rate). Thus, we limit the smallest window length to calculate AI (i.e., epoch length) to be 1-second. Define $\sigma_i(t;H) = \{\sigma_{i1}(t;H), \sigma_{i2}(t;H), \sigma_{i3}(t;H)\}$ to be the standard deviation function for data from a time interval of length H starting at time t. More specifically,

$$\sigma_{im}(t;H) = \sqrt{\frac{\sum_{h=0}^{H-1}\left\{X_{im}(t+h) - \sum_{k=0}^{H-1} X_{im}(t+k)/H\right\}^2}{H}}, \; m = 1,2,3. \tag{4.3}$$

As described in the main text, the $\sigma^2_{im}(t;H)$ are axis- and participant-specific moving variance that characterize the variation of acceleration along each axis in the window of length H starting at t. We then introduce $\bar{\sigma}_i$ as the systematic standard deviation when the device is placed steady (not moving). Formally, for every $t/H \in \mathbb{Z}$,

$$\bar{\sigma}_i = \frac{\sum\limits_{t\in\mathcal{T}_i \text{ and } t/H\in\mathbb{Z}} \sqrt{\frac{1}{3}\left\{\sigma^2_{i1}(t;H) + \sigma^2_{i2}(t;H) + \sigma^2_{i3}(t;H)\right\}}}{|\{t \in \mathcal{T}_i \text{ and } t/H \in \mathbb{Z}\}|}, \tag{4.4}$$

where $\mathcal{T}_i$ stands for collection of time points t during which the device is considered steady. Constraint $t/H \in \mathbb{Z}$ guarantees $\sigma_{im}(t)$s are only computed at time point which is a multiple of H, so that $|\{t \in \mathcal{T}_i \text{ and } t/H \in \mathbb{Z}\}|$ is the number of complete epochs (of length H) in $\mathcal{T}_i$. Usually, $\mathcal{T}_i$ can be specified by the users themselves, and examples include the time period during which the accelerometer is placed on the table. The variances $\bar{\sigma}^2_i$ are the participant-specific systematic noise of acceleration and are calculated as the variance of the observed raw accelerometry data across all three axes in periods of non-movement.

Formally, the Activity Index $\mathrm{AI}_i^{\mathrm{new}}(t;H)$ is given by either

$$\mathrm{AI}_i^{\mathrm{ABS}}(t;H) = \sqrt{(\max\left[\frac{1}{3}\sum_{m=1}^{3}\left\{\sigma_{im}^2(t;H) - \bar{\sigma}_i^2\right\}, 0\right]} \tag{4.5}$$

in absolute scale or

$$\mathrm{AI}_i^{\mathrm{REL}}(t;H) = \sqrt{(\max\left[\frac{1}{3}\left\{\sum_{m=1}^{3}\frac{\sigma_{im}^2(t;H) - \bar{\sigma}_i^2}{\bar{\sigma}_i^2}\right\}, 0\right]}, \tag{4.6}$$

in relative scale. As $\mathrm{AI}_i^{\mathrm{new}}(t;H)$ is defined at every $t \in [1, T_i]$, it is guaranteed to exist for each second.

Note that the $\mathrm{AI}_i^{\mathrm{new}}(t;H)$ is related to, but different from the original AI_0 defined by

$$\mathrm{AI}_i^{0}(t;H) = \max\left[\frac{1}{3}\left\{\sum_{m=1}^{3}\frac{\sigma_{im}(t;H) - \bar{\sigma}'_{im}}{\bar{\sigma}'_{im}}\right\}, 0\right], \tag{4.7}$$

where $\bar{\sigma}'_{im}$ is the systematic noise on each axis computed using the inactive time period, which depends on choosing a threshold C for the distribution of AI_0 in all epochs. The main differences between the original and new AIs are how the systematic noise variance is calculated and how signals from three axes are combined. As will be shown below, compared to the original AI_0, the new AI has a few advantages, including ease for implementation especially in large-scale studies and nice mathematical properties such as additivity and rotational invariance.

4.5.2 Properties of AI

4.5.2.1 Easy implementation

As discussed in Bai et al. [10], the original AI_0 depends on the choice of a threshold C to determine whether a participant was active or not in each second to calculate the systematic noise $\sigma'_{im}(t;H)$. Although $\mathrm{AI}_i^0(t;H)$ was shown not to vary too much with C, this process was tedious and infeasible in large scale studies and might not lead to an AI_0 comparable across different studies. In contrast, Equation4.4 implies that $\bar{\sigma}_i$ is determined using "non-wear periods", which could easily be identified either by existing algorithms, by pilot studies or via participants' self-annotations.

4.5.2.2 Additivity

Additivity of $\mathrm{AI}_i^{\mathrm{new}}(t;H)$ could be formalized as follow. First, let $H = f$ and calculate second-by-second $\mathrm{AI}_i^{\mathrm{new}}(t;f)$. Then, $\mathrm{AI}_i^{\mathrm{new}}(t;f)$s are summed up to generate AI in longer epochs. For example, minute-by-minute AI satisfies

$$\mathrm{AI}_i^{\mathrm{new}}(t;60f) = \sum_{s=0}^{59} \mathrm{AI}_i^{\mathrm{new}}(t+sf;f),$$

where $\mathrm{AI}_i^{\mathrm{new}}(t;60f)$ is only defined at $\{t : t/(60f) \in \mathbb{Z}\}$.

4.5.2.3 Rotational Invariance

In this section we formally prove the rotational invariance of $\mathrm{AI}_i^{\mathrm{new}}(t;H)$. This property is achieved by i) replacing axis-specific $\bar{\sigma}_{i1}$, $\bar{\sigma}_{i2}$ and $\bar{\sigma}_{i3}$ in Equation 4.7

with one single $\bar{\sigma}_i$ as in Equation 4.5 or Equation 4.6 and ii) change standard deviations to variances. Since $\mathrm{AI}_i^{\mathrm{ABS}}(t;H)$ and $\mathrm{AI}_i^{\mathrm{REL}}(t;H)$ are directly proportional, we only give the proof of rotational invariance for $\mathrm{AI}_i^{\mathrm{REL}}(t;H)$ and one can easily follow the same flow to verify such property for $\mathrm{AI}_i^{\mathrm{ABS}}(t;H)$. Let an orthogonal 3×3 matrix $R = \{r_{mm}\}_{3\times 3}$ be the rotation matrix. In another word, rotating $\mathbf{X}_i(t) = \{X_{i1}(t), X_{i2}(t), X_{i3}(t)\}$ with respect to R could be formally written as

$$\mathbf{X}_i^*(t) = \left\{ \sum_{m=1}^{3} r_{1m} X_{im}(t), \sum_{m=1}^{3} r_{2m} X_{im}(t), \sum_{m=1}^{3} r_{3m} X_{im}(t) \right\}, \tag{4.8}$$

while the rotated $\mathbf{X}_i^*(t) = \{X_{i1}^*(t), X_{i2}^*(t), X_{i3}^*(t)\}$ satisfies

$$[X_{i1}^*(t)]^2 + [X_{i2}^*(t)]^2 + [X_{i3}^*(t)]^2 = [X_{i1}(t)]^2 + [X_{i2}(t)]^2 + [X_{i3}(t)]^2,$$

as such rotation does not change the distance from any point to the origin. Further, introduce the mean functions $\mu_{im}(t;N) = \frac{1}{N}\sum_{n=1}^{N} X_{im}(t+n-1)$ and $\mu_{im}^*(t;N) = \frac{1}{N}\sum_{n=1}^{N} X_{im}^*(t+n-1)$ for convenience. It can be verified that the point

$$\{\mu_{i1}(t;N), \mu_{i2}(t;N), \mu_{i3}(t;N)\}$$

is the counterpart of the point

$$\{\mu_{i1}^*(t;N), \mu_{i2}^*(t;N), \mu_{i3}^*(t;N)\}$$

before the rotation R. Since Equation 4.8 holds for any pair of original and rotated points, it guarantees

$$[\mu_{i1}^*(t;N)]^2 + [\mu_{i2}^*(t;N)]^2 + [\mu_{i3}^*(t;N)]^2 = [\mu_{i1}(t;N)]^2 + [\mu_{i2}(t;N)]^2 + [\mu_{i3}(t;N)]^2. \tag{4.9}$$

Finally, together with Equation 4.3 and Equation 4.6, the rotational invariance can be verified as follow

$$\begin{aligned}
\mathrm{AI}_i^{\mathrm{REL}}(t;H) &= \sqrt{\max\left[\frac{1}{3}\left\{\frac{\sigma_{i1}^2(t;H)-\bar{\sigma}_i^2}{\bar{\sigma}_i^2}+\frac{\sigma_{i2}^2(t;H)-\bar{\sigma}_i^2}{\bar{\sigma}_i^2}+\frac{\sigma_{i3}^2(t;H)-\bar{\sigma}_i^2}{\bar{\sigma}_i^2}\right\},0\right]} \\
&= \sqrt{\max\left[\frac{1}{3\bar{\sigma}_i^2}\left\{\sigma_{i1}^2(t;H)+\sigma_{i2}^2(t;H)+\sigma_{i3}^2(t;H)\right\}-1,0\right]} \\
&= \sqrt{\max\left[\frac{1}{3H\bar{\sigma}_i^2}\sum_{m=1}^{3}\sum_{h=0}^{H-1}\left\{X_{im}(t+h)-\mu_{im}(t;H)\right\}^2-1,0\right]} \\
&= \sqrt{\max\left(\frac{1}{3H\bar{\sigma}_i^2}\sum_{h=0}^{H-1}\left[\sum_{m=1}^{3}\left\{X_{im}(t+h)\right\}^2-H\sum_{m=1}^{3}\left\{\mu_{im}(t;H)\right\}^2\right]-1,0\right)} \\
&= \sqrt{\max\left(\frac{1}{3H\bar{\sigma}_i^2}\sum_{h=0}^{H-1}\left[\sum_{m=1}^{3}\left\{X_{im}^*(t+h)\right\}^2-H\sum_{m=1}^{3}\left\{\mu_{im}^*(t;H)\right\}^2\right]-1,0\right)} \\
&= \mathrm{AI}_i^{\mathrm{REL}*}(t;H),
\end{aligned}$$

where $\mathrm{AI}_i^{\mathrm{REL}*}(t;H)$ is the relative scale AI based on rotated data

$$\mathbf{X}_i^*(t) = \left\{X_{i1}^*(t), X_{i2}^*(t), X_{i3}^*(t)\right\}.$$

Chapter 5

A two-stage model for wearable device data

5.1 Background

Recent advances of wearable computing technology have allowed continuous health monitoring in large observational studies and clinical trials. Activity trackers and heart rate monitors are two such devices that are widely used. Activity trackers are used to objectively measure the level and timing of physical activities. They have been used in many scientific studies to supplement or even replace self-reported questionnaires that can be subject to large measurement error and uncontrollable biases [60]. Two comprehensive reviews of such studies were provided by [48] and [94]. Heart rate monitors have an even longer history in health studies (see [83]), because

heart rate is directly related to the physiology of human body and is used extensively to study energy expenditure.

Actiheart (Cambridge Neurotechnology Ltd, Papworth, UK) was one of the earliest wearable devices that combined heart rate and motion sensing [17]. Other integrated multi-sensor trackers have become widely available over the past 5 years both in research and the consumer market [11]. Wearable device data are sampled at a constant rate set by the manufacturer or user. Modern accelerometers can provide densely sampled tri-axial raw accelerometry data, or uni-axial summaries of activity level within various epochs. The raw accelerometry data consist of 3 time series corresponding to the sensor's instantaneous acceleration along 3 orthogonal direction, commonly with 10-100 samples per second on each axis. The raw accelerometry data provide high resolution information on the changes of acceleration, and are standard measurements in modern accelerometry studies. Another type of data is the uni-axial summary data in epochs (in second, minute, etc.), which summarizes information about the intensity of activity from the raw data. Commonly used summary data include various versions of activity count, which are provided by the accelerometer manufacturer, and open-source metrics such as the Activity Index [8] and Mean Amplitude Deviation [89]. This reduced format of data is often preferred in large epidemiological studies, as the size and complexity of the data are reduced while the resulting summaries are considered to be informative enough to describe the daily trajectory of physical activity.

In this paper we will focus on the summary data aggregated at the minute level, which results in 1440 observations per day or more than 10000 per week. The data structure is of the type $Y_{ij}(t)$, where $i(= 1, 2, \ldots, n)$ denotes subject, $j(= 1, 2, \ldots, J_i)$ denotes the day and $t(= 1, 2, \ldots, T_{ij})$ denotes the time of day. The measurement, $Y_{ij}(t)$, is usually a non-negative number with various interpretations, depending on the application. For activity trackers, $Y_{ij}(t)$ is the activity counts or number of steps, which are used as proxies of the activity intensity.

Although such data have been collected in many public health and medical studies, models and analyses are limited to crude summaries, such as the total activity count over a 24-hour period. Despite concerns about the loss of information during the summarization process, only a few analyses utilized the full data. [37, 76, 93] have investigated the association between minute-by-minute activity counts and health covariates, via various methods including functional regression and bivariate smoothing. [55, 56] studied the dynamics of 5-minute energy expenditure level (estimated by accelerometers) using functional principal component analysis and penalized splines. [79] analyzed activity data using principal component analysis to extract the main patterns of variation in activity data. Our goal is to provide a new and systematic methodology to assess the effects of covariates on both the incidence rate and level of physical activity. Although our motivating data are minute-by-minute activity counts (AC), the proposed model can easily be used in other applications where data are densely sampled curves and the outcome is a mixture between a point mass at zero and a

continuous positive distribution.

In Section 5.2, we introduce a two-stage regression model for the transition between active/inactive periods (stage 1) and activity intensity during active periods (stage 2). In Section 5.3, we describe the estimation method and provide the asymptotic properties of the estimators. A simulation study is presented in Section 5.4 to demonstrate the performance of the estimation. Methods are applied in Section 5.5 to the Baltimore Longitudinal Study of Aging (BLSA) to quantify the effect of age, gender and body mass index (BMI) on daily activity trajectories. The paper will conclude with a discussion in Section 5.6.

5.2 Two-stage Model

5.2.1 Notation

We consider the case when the data are collected from a group of subjects ($i = 1, 2, \ldots, n$) during one day ($J_i = 1$) for a fixed number of time units ($T_{ij} = T$). Assume the measurement $Y(t)$ is observed at every time point $t = 1, \ldots, T$. We define the time-dependent binary process $A(t)$, which is an event indicator (e.g. whether the subject was active) at time t. This can be obtained through direct observation or by thresholding $Y(t)$; for example, $A(t) = 1$ if and only if $Y(t) > 0$. If ν is the counting measure on $\{1, 2, \ldots, T\}$ then the total number of events can be written as $\int_0^T A(u)d\nu(u)$, and the total volume of activity proxy from $[0, T]$ is $\int_0^T Y(u)A(u)d\nu(u)$.

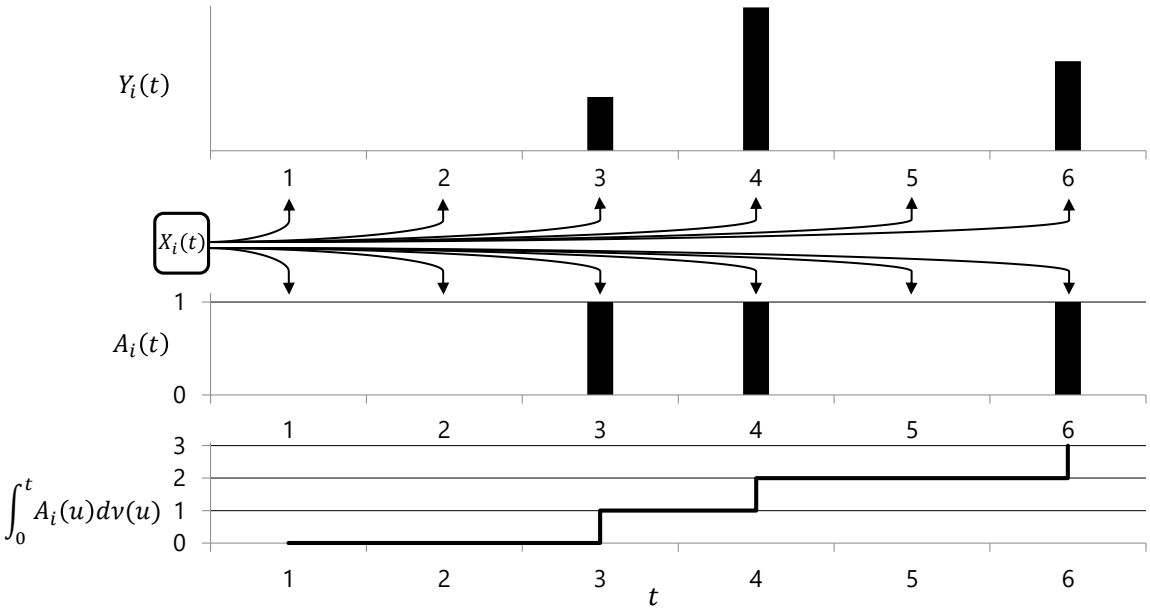


Figure 5.1: Illustration of conceptual model structure. Outcome $Y_i(t)$ is the activity intensity data, observed only at minute $t = 3, 4, 6$ when $A_i(t) = 1$. $\int_0^t A_i(u)dv(u)$ is the total active time up to time t. $X_i(t)$ are covariates which affect both $Y_i(t)$ and $A_i(t)$.

Let $X(t)$ denote the vector of covariates, which can include both time-dependent factors, such as the day time indicator or the average of $Y(t)$ in a past window, and time-independent factors such as baseline age, gender. We assume that the observed data $\{A_i(t), Y_i(t), X_i(t); t = 1, \ldots, T, i = 1, \ldots, n\}$ are i.i.d. replicates of $\{A(t), Y(t), X(t); t = 1, \ldots, T\}$, where i is the subject index.

5.2.2 Model framework

Even though wearable computing data are measured over multiple days, in this paper we will focus on modeling one day of data per subject. This simplifies the

modeling strategy by avoiding the between-day correlation of physical activity within-subject. This can be obtained, for example, by modeling the minute-specific median or mean activity count over multiple activity days of the same subject or by selecting one day of data for each subject (e.g. using the day with the highest total activity count). In our BLSA application result did not change significantly with the choice of data reduction. In future studies we plan to address the problem of having multiple days of data per subject. Figure 5.1 illustrates our conceptual model in the context of activity tracking. For simplicity we only show a 6-minute segment of the data. The top panel contains AC, $Y_i(t)$, with no activity (AC= 0) at minutes 1, 2, and 5. The middle panel contains $A_i(t)$, the process indicating whether AC is greater than 0 at time t. For example, when $t = 3, 4$ or 6, $A_i(t) = 1$ and $\int_0^t A_i(u)d\nu(u)$ increases by 1, indicating the subject was active during that particular time interval. Thus $A_i(t) = I\{Y_i(t) > 0\}$, where $I\{\cdot\}$ is the indicator function. We are interested in quantifying the association between the covariates $X_i(t)$ and the processes $A_i(t)$ and $Y_i(t)$. We further separate $X_i(t)$ into two groups of covariates $Z_i(t)$ and $H_i(t)$, which correspond to covariates with time-invariant effects (those do not vary by time) and covariates with time-varying effects. Separating $Z_i(t)$ and $H_i(t)$ is necessary because the interpretation of their effects is rather different.

In our setting, $A_i(t)$ is observed at every time point $t \in \{1, 2, \ldots, T\}$ and T is relatively large (e.g., $T = 1440$ for one-day of data sampled at the minute level). The notation and structure of the data are similar to those used for continuous marked

processes [68]. However, the modeling strategies we use here are different, because wearable computing data are very large and discretized. Our data could also be analyzed using a standard zero-inflated Poisson count model, but such an approach would be quite involved and relatively hard to implement. Therefore, we choose to present an alternative, two-stage dynamic approach.

5.2.3 Model specification

To assess the effect of covariates $X_i(t)$ on $A_i(t)$ and $Y_i(t)$, we first separate $X_i(t)$ into $Z_i(t)$ and $H_i(t)$ by formally defining $X_i(t) = \{1, Z_i(t)^\intercal, H_i(t)^\intercal\}^\intercal$. We then propose a two-stage model: first stage for $A_i(t)$ and second stage for $Y_i(t)$ conditioning on $A_i(t) = 1$. For simplicity of presentation, notation and interpretation correspond to the case of one vector of time-independent and one vector of time-dependent predictors, but our methodology is not limited by this assumption. First, for each $t = 1, 2, \ldots, T$

$$\text{logit}\mathbb{P}\{A_i(t) = 1 \mid Z_i(t), H_i(t)\} = \beta_0(t) + Z_i(t)^\intercal\beta_1 + H_i(t)^\intercal\beta_2(t). \tag{5.1}$$

At each t, the model assumes that there is a time-varying intercept $\beta_0(t)$ and p_2-dimensional coefficient $\beta_2(t)$, where p_2 is the number of covariates, $H_i(t)$, that have time-dependent effects. There is also a p_1-dimensional structural parameter β_1 across all $t \in [1, T]$, where p_1 is the number of covariates that have time-independent effects. The Stage 1 parameters have standard interpretations for logistic regression

coefficients.

For the second stage, conditioning on $A(t) = 1$, we consider the following semi-parametric model of $Y(t)$

$$\log\{Y_i(t)\} = \gamma_0(t) + Z_i(t)^\intercal \gamma_1 + H_i(t)^\intercal \gamma_2(t) + \epsilon_i(t), \tag{5.2}$$

where $\mathbb{E}\{\epsilon_i(t)\} = 0$, though we make no assumptions about the correlations of $\epsilon_i(t)$ across t within i.

In Model (5.2), we also have the time-varying regression coefficient $\gamma_2(t)$, as well as a vector of structural parameter γ_1 across all t, with dimensions 1 and p_1, respectively. In addition, $\epsilon_i(t)$ is a zero-mean random error with unspecified distribution. The log transformation was applied to $Y_i(t) > 0$ to account for the notoriously heavy right skew of activity counts data [76, 93]. The interpretations of Stage 2 parameters are standard as well. The regression effect parameter of $Z_i(t)$ (or $H_i(t)$) on $Y_i(t)$ is the conditional treatment effect, because our model regresses $Y_i(t)$ by conditioning on $A_i(t) = 1$. More precisely, this is the average effect of $Z_i(t)$ (or $H_i(t)$) on $Y_i(t)$ in the subgroup of the population who are active at time t. Interestingly, the effect of $Z_i(t)$ on $Y_i(t)|A_i(t) = 1$ can also be interpreted as the (composite) treatment effect of $Z_i(t)$ on $Y_i(t) \times A_i(t)$.

Allowing $\beta_2(t)$ and $\gamma_2(t)$ to vary by time is important in practice, as in many applications the time-dependent covariates $H_i(t)$ may have different effects during the day. For example, if $H_i(t)$ is the value of glucose level for subject i at time t, one could expect different association profiles with the subject's physical activity during

day and night.

5.3 Estimation

In this section, we propose an estimating procedure for the unknown parameters in the two-stage model. A reasonable approach could be to use methods that combine fixed effects parameters and flexible spline models for the time-varying parameters. Indeed, in theory, this is a reasonable approach, but in practice we have have been unable to implement it. The problem is that such a model would require fitting one joint model to a very large data set, which makes computations impractical. The problem is that such a model would require fitting one joint model to a very large data set, which makes implementation impractical. Instead, we have used a divide and conquer approach, where we fit minute-specific models to obtain unbiased, but highly variable estimates of the time-invariant and time-variant coefficients and then use smoothing techniques on these coefficients. This approach is much easier to implement and use in practice and we provide both theoretical and numerical evidence that it performs well. Specifically, the estimation starts with treating the structural parameters β_1 and γ_1 as if they were time-varying, i.e. $\beta_1(t)$ and $\gamma_1(t)$. We then estimate $\beta_1(t)$ and $\gamma_1(t)$, along with all the other time-varying parameters, and obtain the estimated $\hat{\beta}_1(t)$ and $\hat{\gamma}_1(t)$, for all $t = 1, 2, \ldots, T$. To simplify notations we define the p-dimensional ($p = p_1 + p_2 + 1$) parameters $\beta(t) = \{\beta_0(t), \beta_1(t)^\intercal, \beta_2(t)^\intercal\}^\intercal$

and $\gamma(t) = (\gamma_0(t), \gamma_1(t)^\intercal, \gamma_2(t)^\intercal)^\intercal$, with the covariates $X_i(t) = \{1, Z_i^\intercal, H_i(t)^\intercal\}^\intercal$ at each t.

The Stage 1 parameter $\beta(t)$ can be estimated by solving the $p \times 1$ score equations $U_{1t}\{\beta(t)\} = 0$, which are defined as

$$U_{1t}\{\beta(t)\} = \frac{1}{n}\sum_{i=1}^{n} U_{1it}\{\beta(t)\} = \frac{1}{n}\sum_{i=1}^{n} X_i(t)\left[A_i(t) - p\{X_i(t); \beta(t)\}\right], \tag{5.3}$$

and $p\{X_i(t); \beta(t)\}$ is the probability of being active at t given the covariates $X_i(t)$, that is, $p\left\{X_i(t); \beta(t)\right\} = \exp\{X_i(t)^\intercal\beta(t)\}/\left[1 + \exp\{X_i(t)^\intercal\beta(t)\}\right]$. We denote the solution of $U_{1t}\{\beta(t)\} = 0$ as $\hat{\beta}(t) = \{\hat{\beta}_0(t), \hat{\beta}_1(t)^\intercal, \hat{\beta}_2(t)^\intercal\}^\intercal$. It is easy to show that for each t, the random vector $\sqrt{n}\{\hat{\beta}(t) - \beta(t)\}$ converges weakly to a multivariate normal random vector as $n \to \infty$.

To estimate the Stage 2 parameter $\gamma(t)$, we use the equality

$$\mathbb{E}[\log\{Y(t)\} - X(t)^\intercal\gamma(t) \mid A(t) = 1] = 0$$

to construct the following $p \times 1$ estimation functions,

$$U_{2t}\{\gamma(t)\} = \frac{1}{n}\sum_{i=1}^{n} U_{2it}\{\gamma(t)\} = \frac{1}{n}\sum_{i=1}^{n} X_i(t)A_i(t)\left[\log\{Y_i(t)\} - X_i(t)^\intercal\gamma(t)\right]. \tag{5.4}$$

We can solve the estimating equation $U_{2t}\{\gamma(t)\} = 0$ to obtain the estimated $\hat{\gamma}(t)$. The estimating equations (5.4) are inspired by the expected value of the "total magnitude of activity over time", or $\mathbb{E}[Y_i(t) \times A_i(t) \mid X_i(t)] = \exp\{X_i(t)^\intercal\gamma(t)\}p\{X_i(t); \beta(t)\}$. This quantity combines information from both stages and shows how the total activity is accumulated.

Recall that β_1 and γ_1 were estimated as time-varying; we now propose two methods to combine information across $t(= 1, 2, \ldots, T)$ to obtain the estimates of β_1 and γ_1. Define

$$\begin{aligned} & U_i(\beta_1, \gamma_1) \\ = \; & [U_{1i1}\{\beta_0(1), \beta_1, \beta_2(1)\}^\intercal, U_{1i2}\{\beta_0(2), \beta_1, \beta_2(2)\}^\intercal, \ldots, U_{1iT}\{\beta_0(T), \beta_1, \beta_2(T)\}^\intercal, \\ & U_{2i1}\{\gamma_0(1), \gamma_1, \gamma_2(1)\}^\intercal, U_{2i2}\{\gamma_0(2), \gamma_1, \gamma_2(2)\}^\intercal, \ldots, U_{2iT}\{\gamma_0(T), \gamma_1, \gamma_2(T)\}^\intercal]^\intercal, \end{aligned}$$

where $U_{1it}\{\cdot\}$ and $U_{2it}\{\cdot\}$ follow a similar formulation as Equations (5.3) and (5.4) without treating β_1 and γ_1 as time-varying. We then plug in the estimated $\hat{\beta}_0(t)$, $\hat{\beta}_2(t)$, $\hat{\gamma}_0(t)$ and $\hat{\gamma}_2(t)$ to form the profile estimation function

$$\begin{aligned} & \tilde{U}_i(\beta_1, \gamma_1) \\ = \; & \Big[U_{1i1}\{\hat{\beta}_0(1), \beta_1, \hat{\beta}_2(1)\}^\intercal, U_{1i2}\{\hat{\beta}_0(2), \beta_1, \hat{\beta}_2(2)\}^\intercal, \ldots, U_{1iT}\{\hat{\beta}_0(T), \beta_1, \hat{\beta}_2(T)\}^\intercal, \\ & U_{2i1}\{\hat{\gamma}_0(1), \gamma_1, \hat{\gamma}_2(1)\}^\intercal, U_{2i2}\{\hat{\gamma}_0(2), \gamma_1, \hat{\gamma}_2(2)\}^\intercal, \ldots, U_{2iT}\{\hat{\gamma}_0(T), \gamma_1, \hat{\gamma}_2(T)\}^\intercal]^\intercal. \end{aligned}$$

The first method is to consider a weighted average of the profile estimating functions. Specifically, let W_1 be a $2p_1 \times 2Tp$ weight matrix and we solve the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} W_1 \tilde{U}_i(\beta_1, \gamma_1) = 0,$$

to obtain the estimator $(\hat{\beta}_1^A, \hat{\gamma}_1^A)$. This approach constructs an estimator of the $2p_1$-dimensional parameter $(\beta_1^\intercal, \gamma_1^\intercal)^\intercal$ by augmenting data information from the $2Tp$ estimation equations. Denoting the true coefficients by β_1^0 and γ_1^0, it is shown in Web

Appendix A that under regularity conditions, $\sqrt{n}\{(\hat{\beta}_1^A - \beta_1^0)^\intercal, (\hat{\gamma}_1^A - \gamma_1^0)^\intercal\}^\intercal$ converges in distribution to a zero mean normal random vector.

The second method combines information across time points, using the approach of the Generalized Method of Moments (GMM) [41], and estimate the regression parameters by minimizing a "general distance" from the estimating functions to zero. Define the weight matrix W_2 with $2Tp \times 2Tp$ dimension, the estimators $(\hat{\beta}_1^B, \hat{\gamma}_1^B)$ is computed via the minimization procedure

$$(\hat{\beta}_1^B, \hat{\gamma}_1^B) = \underset{(\beta_1, \gamma_1)}{\arg\min} \left\{ \sum_{i=1}^{n} \tilde{U}_i(\beta_1, \gamma_1) \right\}^\intercal W_2 \left\{ \sum_{i=1}^{n} \tilde{U}_i(\beta_1, \gamma_1) \right\}. \tag{5.5}$$

We show in Section 5.7 that $\sqrt{n}\{(\hat{\beta}_1^B - \beta_1^0)^\intercal, (\hat{\gamma}_1^B - \gamma_1^0)^\intercal\}^\intercal$ converges in distribution to a zero mean normal random vector.

In practice, the weight matrices W_1 and W_2 are pre-specified by investigators, and the estimation efficiency may vary when using different weights. According to the GMM theory, the matrix $W_2 = \Sigma^{-1}$, where Σ is defined in Section 5.7, yields the most asymptotically efficient estimator in the class of all GMM estimators. Moreover, if we set $W_1 = D(\beta_1^0, \gamma_1^0)^\intercal W_2$, the two approaches will be asymptotically equivalent, where $D(\beta_1^0, \gamma_1^0)$ is defined in Section 5.7. However, even if the weights are not optimal, combining information across time points is expected to substantially improve the performance of estimators compared to using a single time point.

Note that both methods are proposed to address the over-identification of the estimating equations, by borrowing information across t to improve the estimation efficiency of structural parameters β_1 and γ_1. However, when T is large (i.e. $T = 1440$

for minute-by-minute activity count), the direct application of GMM could be computationally challenging when minimizing a function with $2(1+p_2)T+2p_1$ parameters. To solve this problem, we substitute the unknown parameter ($\beta_0(t)$,$\gamma_0(t)$, $\beta_2(t)$,$\gamma_2(t)$) in the estimating equations with their estimate from regression models at each time point t, and obtain $2pT$ estimating equations with $2p_1$ parameters. GMM is then used to combine the estimating estimating equations. In practice, it may be difficult to estimate or calculate the optimal weight matrix $W_2 = \Sigma^{-1}$ (e.g., T is too large and inversion of Σ is computationally expensive). In such situations we can consider the other method, which directly combines the estimating equations using a pre-specified weight matrix W_1. Note that, in general, for arbitrary W_1 and W_2, either approach could be more efficient; however, with $W_1 = D(\beta_1^0, \gamma_1^0)^\intercal W_2$, the two approaches are asymptotically equivalent.

5.4 Simulation Study

We demonstrate the performance of our method via a simulation study, with the data generating mechanism inspired by the observed accelerometry data. The simulation was conducted in 4 scenarios, each with different sample size ($N = 300$ or 600) and/or time span ($T = 300$ or 600).

For each scenario we simulated $R = 300$ datasets starting with $A_i(t)$ $(i = 1, \ldots, N)$

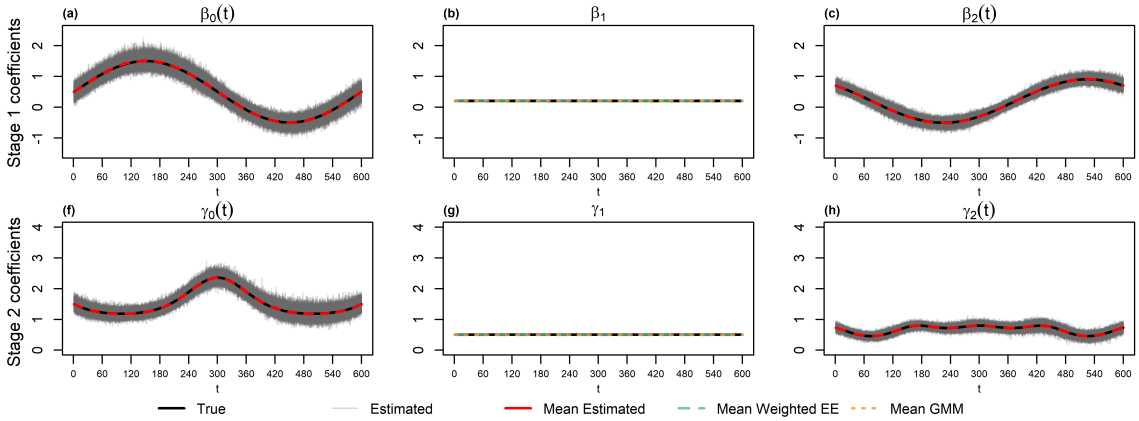


Figure 5.2: Comparison of true coefficient curves (black), mean estimated coefficient curves for all time-varying coefficients(red), two types of estimates (green for weighted estimation equation and orange for GMM) for the structural parameters, and estimated coefficient curves (gray) when $N = 600$, $T = 600$.

according to the model

$$\text{logit}\mathbb{P}[A_i(t) = 1 \mid Z_i, H_i(t)] = \beta_0(t) + Z_i\beta_1 + H_i(t)\beta_2(t). \tag{5.6}$$

For each $t = 1, 2, \ldots, T$, we let $\beta_0(t) = \sin(2\pi t/T) + 0.5$, $\beta_1 = 0.2$, $\beta_2(t) = 0.2 - \frac{1}{2}\{\sin(2\pi t/T) - \cos(2\pi t/T)\}$ (black solid curves in Figure 5.2). The scalar predictors in Model (5.6) are simulated for each i and t from $Z_i \sim N(1/2, 1/2)$ and $H_i(t) = 1.5\sin(t/50 + \delta_i) + Q + 0.5$, where $Q \sim N(0, 0.2)$ and $\delta_i \sim \text{unif}\{1, 2, \ldots, 10\}$. $A_i(t)$'s were correlated binary outcomes with the marginal model $\mathbb{P}[A_i(t) = 1 \mid Z_i, H_i(t)] = \text{logit}^{-1}\{\beta_0(t) + Z_i\beta_1 + H_i(t)\beta_2(t)\}$ while the T-dimensional correlation matrix $\mathbf{R} = \{r_{lm}\}_{T\times T}$, where $r_{lm} = 0.5^{|l-m|I\{|l-m|<10\}}$. The non-zero observations are generated from the model

$$\log[Y_i(t) \mid A_i(t) = 1, Z_i, H_i(t)] = \gamma_0(t) + Z_i\gamma_1 + H_i(t)\gamma_2(t) + \epsilon_i(t). \tag{5.7}$$

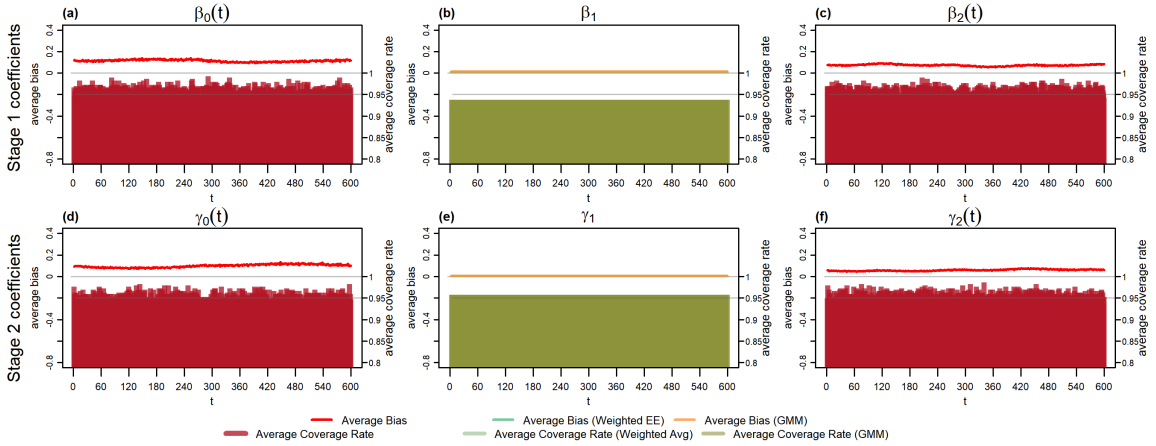


Figure 5.3: Average bias for all time-varying parameters (red), average bias for the structural parameters via weighted estimation equation (green) and via GMM (orange). Average coverage for all time-varying parameters (red), average bias for the structural parameters via weighted estimation equation (green) and via GMM (brown). $N = 600$, $T = 600$.

The Stage 2 parameters are given by $\gamma_0(t) = \sqrt{\exp\{\sin(3\pi t/T - \pi)\}/2 + 1}$, $\gamma_1 = 0.5$ and $\gamma_2(t) = 3\phi_0\{\sin(4\pi t/T - \pi) + \arctan(2\pi t/T - \pi)/2\}$ (black curves in panels (f), (g) and (h) in Figure 5.2; $\phi_0(\cdot)$ is the probability density function of the distribution $N(0, 1.5)$). Data are then generated as $Y_i(t) = \exp\{\gamma_0(t) + Z_i\gamma_1 + H_i(t)\gamma_2(t) + \epsilon_i(t)\}$, where for each i, $\epsilon_i(t)$ follows a multivariate normal distribution with mean $\mathbf{0}$ and the same T-dimensional correlation matrix $\mathbf{R}$ as in Stage 1.

To quantify estimation accuracy, we calculated the average absolute bias and mean integrated square error (MISE) at every t. If $\hat{\eta}_r(t)$ is the estimator of the true curve $\eta(t)$ using the rth simulated data, then the average absolute bias for this coefficient at time t is defined as $\frac{1}{R}\sum_{r=1}^{R}|\eta(t) - \hat{\eta}_r(t)|$, and the mean integrated square error $\text{MISE} = \frac{1}{TR}\sum_{r=1}^{R}\sum_{t=1}^{T}\{\eta(t) - \hat{\eta}_r(t)\}^2$ (note that the discrete version of MISE is the

average MSE across t). We also calculated the coverage rate of the bootstrap (non-parametric bootstrap of subjects repeated for 300 times) pointwise 95% confidence intervals (obtained via estimated value plus/minus 1.96 times bootstrapped standard error of the estimated value) at every t. Figure 5.3 provides a visual comparison between the average absolute bias (solid curves) and the average coverage rate (vertical bars) in the case when $N = 600$, $T = 600$. The average absolute bias for all coefficients remains small although varying slightly with t. The coverage rate of all coefficients in both stages is around 0.95. For the estimation of β_1 and γ_1, the weighted estimation equation and GMM have inseparable performance according to the plot, both in terms of bias and confidence interval coverage rate. Figure 5.2 provides a comparison between the true coefficient curves (shown in black) and three different types of mean estimated coefficient curves ($N = 600$, $T = 600$). Time-varying parameters are shown in red, and time-invariant parameters via weighted estimation equation and GMM are shown in green and orange, respectively. The 300 estimated coefficient curves for all time-varying and -invariant parameters are shown in gray. The true and mean estimated coefficient curves are in good agreement for all coefficients and all time points t. The estimated coefficient curves exhibit a roughly constant variability across t around the true $\beta_0(t)$ and $\beta_2(t)$.

Table 5.1 provides MISE and average coverage rate across t, while using either weighed estimation equation or GMM to estimate the time-invariant parameters. When holding T constant and increasing the sample size N from 300 to 600, MISE

Table 5.1: Mean integrated squared error MISE $= \frac{1}{TR}\sum_{r=1}^{R}\left[\sum_{t=1}^{T}\{\eta(t)-\hat{\eta}_r(t)\}^2\right]$ and average coverage rate across t using weighted estimation equation and GMM method in 4 different scenario.

			$\beta_0(t)$	β_1	$\beta_2(t)$	$\gamma_0(t)$	γ_1	$\gamma_2(t)$
$N=300,\ T=300$	W/t EE	MISE	3.9×10^{-2}	4.6×10^{-4}	1.7×10^{-2}	1.3×10^{-2}	1.6×10^{-4}	5.2×10^{-3}
		Coverage	95.6%	96.0%	95.5%	94.5%	93.7%	94.7%
	GMM	MISE	3.9×10^{-2}	6.5×10^{-4}	1.7×10^{-2}	1.3×10^{-2}	3.2×10^{-4}	5.2×10^{-3}
		Coverage	95.6%	92.7%	95.5%	94.5%	94.3%	94.7%
			$\beta_0(t)$	β_1	$\beta_2(t)$	$\gamma_0(t)$	γ_1	$\gamma_2(t)$
$N=300,\ T=600$	W/t EE	MISE	4.1×10^{-2}	2.5×10^{-4}	1.7×10^{-2}	1.4×10^{-2}	7.7×10^{-5}	5.0×10^{-3}
		Coverage	95.5%	94.3%	95.5%	94.5%	97.0%	94.6%
	GMM	MISE	4.1×10^{-2}	3.9×10^{-4}	1.7×10^{-2}	1.4×10^{-2}	1.5×10^{-4}	5.0×10^{-3}
		Coverage	95.5%	92.3%	95.5%	94.5%	95.0%	94.6%
			$\beta_0(t)$	β_1	$\beta_2(t)$	$\gamma_0(t)$	γ_1	$\gamma_2(t)$
$N=600,\ T=300$	W/t EE	MISE	2.0×10^{-2}	2.0×10^{-4}	8.6×10^{-3}	7.0×10^{-3}	7.6×10^{-5}	2.6×10^{-3}
		Coverage	95.4%	97.0%	95.1%	94.7%	96.3%	94.7%
	GMM	MISE	2.0×10^{-2}	2.5×10^{-4}	8.6×10^{-3}	7.0×10^{-3}	1.4×10^{-4}	2.6×10^{-3}
		Coverage	95.4%	96.0%	95.1%	94.7%	96.3%	94.7%
			$\beta_0(t)$	β_1	$\beta_2(t)$	$\gamma_0(t)$	γ_1	$\gamma_2(t)$
$N=600,\ T=600$	W/t EE	MISE	2.1×10^{-2}	1.4×10^{-4}	8.4×10^{-3}	7.1×10^{-2}	5.3×10^{-5}	9.0×10^{-5}
		Coverage	95.2%	95.0%	95.2%	94.7%	93.7%	94.7%
	GMM	MISE	2.1×10^{-2}	1.8×10^{-4}	8.4×10^{-3}	7.1×10^{-2}	9.0×10^{-5}	9.0×10^{-5}
		Coverage	95.2%	93.3%	95.2%	94.7%	95.7%	94.7%

drops noticeably (by about a half). When holding N constant and increasing T from 300 to 600, MISE of time-varying coefficients remains about the same but that of time-invariant coefficients decreases substantially. On the other hand, the coverage rate always remains close to 0.95 in all 4 of our scenarios. The weighted estimation equation method exhibited smaller bias but better coverage rate than GMM when estimating β_1 and γ_1.

5.5 Application

The Baltimore Longitudinal Study of Aging (BLSA), funded and operated by the National Institute of Aging (NIA), is a study of normative human aging established in 1958 and continuing to this day. Briefly, BLSA continuously enrolls community volunteers as subjects who pass a series of health and functional evaluations. All participants are followed for life and visited every one to four years depending on age. More detailed description of the study could be found in [34]. The sample for the current study consists of men and women who underwent a physical examination, health history, and comprehensive energy expenditure testing during their visit between September 2007 and August 2015. The participants were admitted to the Clinical Research Branch unit of the National Institute on Aging for 3 days of testing. Height and weight were assessed in light clothing using a stadiometer and calibrated scale, respectively. Date of birth (age) was derived from a health history interview

conducted by trained technicians. On the last day of the visit, participants were asked to wear the Actiheart (CamNtech Ltd, Papworth, UK) device attached on the chest using two standard electrodes. The device collected both heart rate and activity counts in one-minute epochs for the subsequent 7 days in the free-living environment, and was returned to NIA via FedEx.

The data used in this paper were collected from 878 participants who had at least 3 full days of monitoring of their physical activity. More specifically, the data for each subject are minute-by-minute time series of Activity Count (AC). More intense activity tends to correspond to higher AC, at least as measured at the chest level. An AC value of 0 count corresponds to no activity detected. We start by denoting the observed AC of subject i $(= 1, 2, \ldots, I)$ during the t^{th} minute of day d $(= 1, 2, \ldots, D_i)$ by $Y^0_{id}(t)$, $t = 1, 2, \ldots, 1440$. In this data analysis, we only included the most active day (in terms of daily average AC) during each visit for each subject. Specifically, we defined $Y_i(t) = Y^0_{id^*}(t)$ where $d^* = \underset{d\in\{1,2,\ldots,D_i\}}{\text{argmax}} [\sum_{t=0}^{T} Y^0_{id}(t)/T]$ for each subject i at time t. To these data we apply the two-stage model described in Section 5.2 and investigate the association between demographic factors and level of activity in the past one hour and current level of activity at each time of day. The demographic factors in our analysis were time-independent: age (centered at 70), gender (male as 1) and BMI (centered at 27). The minute by minute heart rate $H_i(t)$ (centered at 72) was included as time-dependent covariates. These covariates are basic demographic variables which might only partially account for the population heterogeneity of baseline

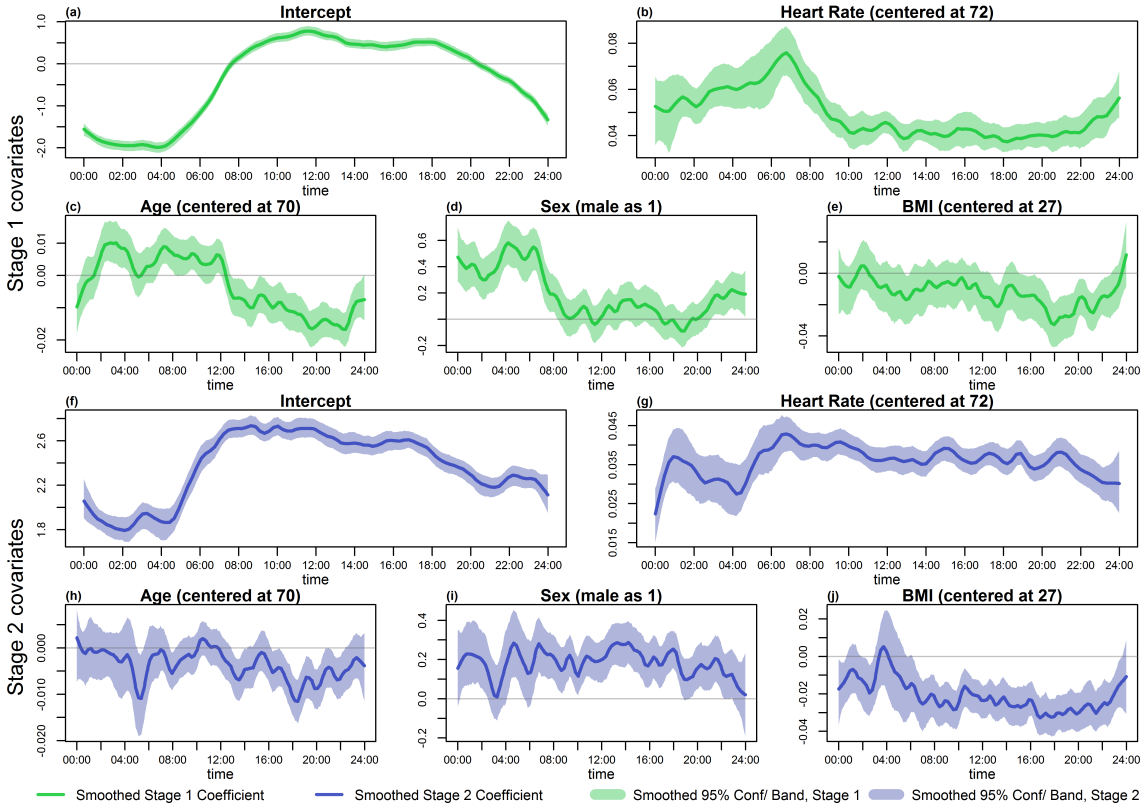


Figure 5.4: Estimated covariate coefficients of Stage 1 and 2 models. Green curves depict LOWESS smoothed coefficients of the Stage 1 model, while light green shadows stand for pointwise bootstrapped 95% confidence interval. Blue curves and shadows follow the same illustration, but for Stage 2 model.

physical activity. There are many other factors that might affect the baseline physical activity, such as the fitness level, seasonality, geographical location and so on. We only included basic demographics and heart rate to keep this section compact and focus on illustrating application of the two-stage model. We start with modeling all coefficients as time-varying.

Figure 5.4 presents the estimated coefficients of the two-stage model applied to BLSA. The first two rows of the figure depict the coefficient curves at each time t for

the Stage 1 model, while the next two rows display the estimated coefficients for the Stage 2 model. The darker green and blue curves provide a LOWESS smoother for the estimated coefficients. The light green and blue shades are pointwise bootstrap 95% confidence bands, which were obtained from LOWESS smoothed coefficients estimated using bootstrapped samples. Figure 5.4(a) and (f) illustrate the pattern of the intercept for Stage 1 and 2 models, respectively. Figure 5.4(a) indicates that, on average, a 70-year old woman with a BMI of 27 and heart rate of 72 is less likely to be active during the night (6PM-8AM), but roughly maintains a 60% probability of being active during the daytime from 8AM to 8PM. Figure 5.4(f) complements Figure 4(a), suggesting that if the same group of women are active during daytime then the intensity of their activity declines, on average, starting from 12PM. Figures 5.4(b) and 5.4(g) display the time-varying Stage 1 and Stage 2 coefficients for the minute by minute heart rate. Both coefficients indicate a positive association for the odds of moving and for activity count at every time t. Figure 5.4(b) indicates that one unit of increased heart rate is associated to greater odds of being active at night and earlier morning compared to day time. However, the Stage 2 coefficient in Figure 5.4(g) exhibits a more consistent day/night pattern. Figures 5.4(c) and 5.4(h) display the effect of age as a function of time during the day for the Stage 1 and Stage 2 models. Together, they indicate that older individuals are more likely to move during the night and early morning periods, though they are less likely to move after 12PM. If they do engage in physical activity then their activity level is, on average, lower

in the afternoon and evening (12PM to 22PM). There is also a large increase of the negative effect around 6PM in panel 5.4(h), which suggests that the largest reduction in activity level for older individuals happens in late afternoon. Results indicate that physical activity interventions may be more effective at increasing overall daily activity if they are targeted towards the second part of the day. Figure 5.4(d) and (i) illustrate the gender effect. They indicate that the odds of being active is roughly the same for men and women during the day but it is substantially higher for men during the night. In addition, when comparing active men and active women, the level of activity is consistently higher for men throughout day. Because the sex effect seems to depend only on the day or night period, we further considered two time-invariant sex effects, one for day and one for night time. Figure 5.5 displays the various sex effect estimates. Panel (a) and (d) show the time-varying gender effect as in Figure 5.4. The weighted estimation equation (panels (b) and (e)) and the GMM method (panels (c) and (f)) both yielded a negative stage 1 effect at night, roughly zero stage 1 effect at daytime and a significant stage 2 effect throughout the day; these results are consistent with our observation from Figure 5.5(a) and (d). As more information across t was taken into account to estimate the time-invariant sex effect, we chose to leave it in the final model.

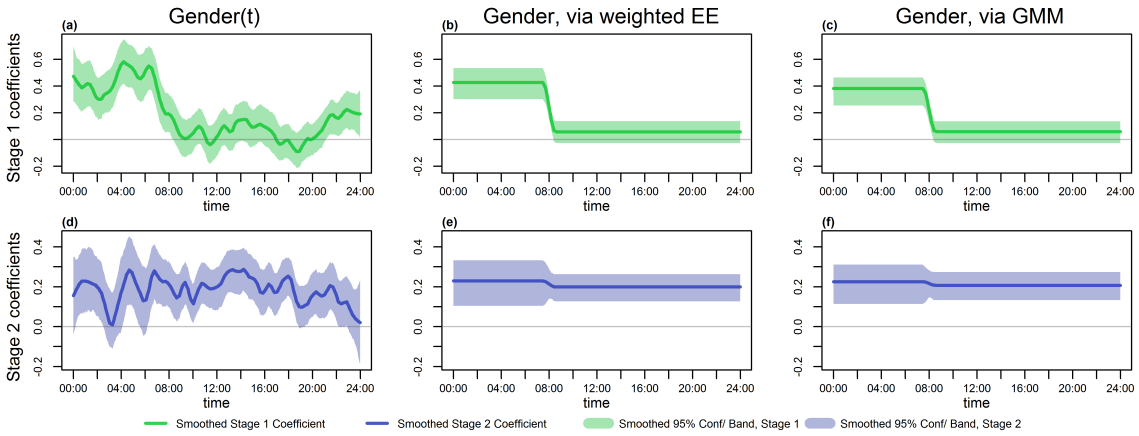


Figure 5.5: Comparison of estimated coefficients of gender: as time-varying effect in (a) and (d); as time-invariant coefficient estimated via weighted estimation equation in (b) and (e); as time-invariant coefficient estimated via GMM in (c) and (f). The time-invariant coefficients were fitted separately for the nighttime (0AM-8AM) and daytime (otherwise). Green curves depict LOWESS smoothed coefficients of the Stage 1 model, while light green shadows stand for pointwise bootstrapped 95% confidence interval. Blue curves and shadows follow the same illustration, but for Stage 2 model.

5.6 Discussion

We introduced a two-stage model for a general type subject-specific, dense time series data with excess zeros. Such data are typical in studies using wearable devices. The model describes the effect of covariates on the occurrence of events and, conditional on the fact that the event occurred, it describes the effect of covariates on the magnitude of response. A logistic regression model was introduced to account for the first stage of the model, and a log-linear model was used for the second stage of the model. While both models are relatively well known, the novel combination, the application to high density wearable computing data, and the extension of zero-inflated models to time series make our approach novel. Most importantly, such approaches

are absolutely crucial to uncover the type of findings we have shown in Section 5.5, for generating hypothesis, and for suggesting simpler, easier to understand models.

Our scientific findings are both striking and reasonable. We contend that this is the first time when the effect of covariates' is separated according to probability of being active and the level of activity. Moreover, going from qualitative statements about possible associations that seem reasonable to a reader to quantifying these associations is often the difference between research and hearsay. We have shown that some covariates affect the odds of being active while others only affect the level of activity. Such separation of the covariate effect is important, as researchers may start to disentangle the "will" and "ability" of being active. Note, for example, that in-lab experiments often have a very difficult time distinguishing between two equally able individuals who have a very different levels of motivation to be active. In the lab both would perform equally well. In the natural living environment things can be completely different. Ultimately, this could provide valuable information in terms of targeting interventions and understanding the complex nature of interactions between physiological and psychological determinants of activity and health. Our work has laid a foundation for more systematic ways of analyzing wearable computing data in the future.

Our method has several limitations that could be further addressed in future studies. First, we did not assume independence across t because the two-stage model is a marginal model which treats correlations across t as nuisance parameters. In the

future, it would be of great interest to develop statistical techniques that could take advantage of, for example, local correlations from repeated measurements (across t) to improve estimation efficiency. Second, we only used one day of accelerometry data within each multiple-day visit. An important extension would be to develop methods that take into account the correlation structure of physical activity profiles across several days within each visit. Third, we estimated each time-varying coefficient at every t and smoothed them afterwards. An alternative way could be to introduce smoothing directly during the estimation. For example, a set of q-dimensional B-spline basis functions might be used to effectively reduce the number of parameters for one covariate from T to q, where $q \ll T$ is the number of parameters of the spline. Moreover, one could use periodic splines to better capture the periodic behavior of coefficients and ensure that the same value is estimated at the start (0AM) and end (24PM) of the day. This may further reduce the standard error of time-varying coefficients close to the middle of the night.

5.7 Supplementary Material

5.7.1 Large-sample property of the estimators

To simplify the discussion, we assume the true regression parameters $\beta^0(t) = \{\beta_0^0(t)^\intercal, \beta_1^0, \beta_2^0(t)^\intercal\}^\intercal$ and $\gamma^0(t) = \{\gamma_0^0(t)^\intercal, \gamma_1^0, \gamma_2^0(t)^\intercal\}^\intercal$ lie in a compact set. We first

establish the asymptotic properties of the estimators $\hat{\beta}(t)$ and $\hat{\gamma}(t)$. Since

$$\begin{aligned}\sqrt{n}\{\hat{\beta}(t)-\beta^0(t)\} &= n^{-1/2}\sum_{i=1}^{n}\Gamma_1(t)^{-1}X_i(t)\{A_i(t)-p\{X_i(t);\beta_0(t)\}\}+o_p(1)\\ &\equiv n^{-1/2}\sum_{i=1}^{n}\phi_i(t)+o_p(1)\end{aligned}$$

where $\Gamma_1(t)=\mathbb{E}\left[\exp\{X(t)^\intercal\beta^0(t)\}/\{1+\exp\{X(t)^\intercal\beta^0(t)\}\}^2X(t)X(t)^\intercal\right]$. Thus we have $\sqrt{n}\{\hat{\beta}_0(t)-\beta_0^0(t),\hat{\beta}_2(t)-\beta_2^0(t)\}=n^{-1/2}\sum_{i=1}^{n}a_i(t)+o_p(1)$, where $a_i(t)$, the vector of first and last p_2 elements of $\phi_i(t)$, are i.i.d. mean zero random vectors. Similarly, since

$$\begin{aligned}\sqrt{n}\{\hat{\gamma}(t)-\gamma^0(t)\} &= n^{-1/2}\sum_{i=1}^{n}\Gamma_2(t)^{-1}X_i(t)A_i(t)[\log\{Y_i(t)\}-X_i(t)^\intercal\gamma^0(t)]+o_p(1)\\ &\equiv n^{-1/2}\sum_{i=1}^{n}\psi_i(t)+o_p(1)\end{aligned}$$

where $\Gamma_2(t)=E[A(t)X(t)X(t)^\intercal]$. Thus we have $\sqrt{n}\{\hat{\gamma}_0(t)-\gamma_0^0(t),\hat{\gamma}_2(t)-\gamma_2^0(t)\}=n^{-1/2}\sum_{i=1}^{n}b_i(t)+o_p(1)$, where $b_i(t)$, the vector of first and last p_2 elements of $\psi_i(t)$, are i.i.d. mean zero random vectors.

Then we define profile estimating functions

$$\tilde{U}_{1t}(\beta_1)=U_{1t}(\hat{\beta}_0(t),\beta_1,\hat{\beta}_2(t)),$$

$$\tilde{U}_{2t}(\gamma_1)=U_{2t}(\hat{\gamma}_0(t),\gamma_1,\hat{\gamma}_2(t)),$$

$$\tilde{U}_1(\beta_1)=\{\tilde{U}_{11}(\beta_1)^\intercal,\tilde{U}_{12}(\beta_1)^\intercal,\ldots,\tilde{U}_{1T}(\beta_1)^\intercal\}^\intercal,$$

$$\tilde{U}_2(\gamma_1)=\{\tilde{U}_{21}(\gamma_1)^\intercal,\tilde{U}_{22}(\gamma_1)^\intercal,\ldots,\tilde{U}_{2T}(\gamma_1)^\intercal\}^\intercal$$

and

$$\tilde{U}(\beta_1,\gamma_1)=\{\tilde{U}_1(\beta_1)^\intercal,\tilde{U}_2(\gamma_1)^\intercal\}^\intercal.$$

By defining the derivatives with respect to $(\beta_0(t), \beta_2(t))$,

$$B_{1t}(\beta_0^0(t), \beta_2^0(t)) = \partial \mathbb{E}\{U_{1t}(\beta_0(t), \beta_2(t), \beta_1^0)\}/\partial(\beta_0(t), \beta_2(t))|_{(\beta_0^0(t), \beta_2^0(t))},$$

and

$$B_{2t}(\gamma_0^0(t), \gamma_2^0(t)) = \partial \mathbb{E}\{U_{2t}(\gamma_0(t), \gamma_2(t), \beta_1^0)\}/\partial(\gamma_0(t), \gamma_2(t))|_{(\gamma_0^0(t), \gamma_2^0(t))},$$

we have

$$\begin{aligned}\sqrt{n}\tilde{U}_{1t}(\beta_1^0) =& n^{-1/2}\sum_{i=1}^{n}\left\{U_{1it}(\beta_0^0(t), \beta_1^0, \beta_2^0(t)) + B_{1t}(\beta_0^0(t), \beta_2^0(t))a_i(t)\right\} + o_p(1)\\ \equiv& n^{-1/2}\sum_{i=1}^{n} c_{1i}(t) + o_p(1),\end{aligned}$$

and

$$\begin{aligned}\sqrt{n}\tilde{U}_{2t}(\gamma_1^0) =& n^{-1/2}\sum_{i=1}^{n}\left\{U_{2it}(\gamma_0^0(t), \gamma_2^0(t), \gamma_1^0) + B_{2t}(\gamma_0^0(t), \gamma_2^0(t))b_i(t)\right\} + o_p(1)\\ \equiv& n^{-1/2}\sum_{i=1}^{n} c_{2i}(t) + o_p(1).\end{aligned}$$

Thus, we have $\sqrt{n}\tilde{U}(\beta_1^0, \gamma_1^0) = n^{-1/2}\sum_{i=1}^{n} C_i + o_p(1)$, where

$$C_i = \{c_{1i}(1)^\intercal, c_{1i}(2)^\intercal, \ldots, c_{1i}(T)^\intercal, c_{2i}(1)^\intercal, c_{2i}(2)^\intercal, \ldots, c_{2i}(T)^\intercal\}^\intercal.$$

To study the asymptotic properties of the estimators $(\hat{\beta}_1^A, \hat{\gamma}_1^A)$ and $(\hat{\beta}_1^B, \hat{\gamma}_1^B)$, for every t, we define the partial derivatives of the estimating functions with respect to β_1 and γ_1 as $\tilde{D}_{1t}(\beta_1) = \partial\tilde{U}_{1t}(\beta_1)/\partial\beta_1$, $\tilde{D}_{2t}(\gamma_1) = \partial\tilde{U}_{2t}(\gamma_1)/\partial\gamma_1$, $\tilde{D}_1(\beta_1) = \{\tilde{D}_{11}(\beta_1)^\intercal, \ldots, \tilde{D}_{1T}(\beta_1)^\intercal\}^\intercal$ and $\tilde{D}_2(\gamma_1) = \{\tilde{D}_{21}(\gamma_1)^\intercal, \ldots, \tilde{D}_{2T}(\gamma_1)^\intercal\}^\intercal$. We further define their limits as $D_{1t}(\beta_1) = \partial\mathbb{E}\{U_{1t}(\beta_0(t), \beta_1, \beta_2(t))\}/\partial\beta_1$ and $D_{2t}(\gamma_1) = \partial\mathbb{E}\{U_{2t}(\gamma_0(t), \gamma_1, \gamma_2(t))\}/\partial\gamma_1$, which reduced to $D_1(\beta_1) = \{D_{11}(\beta_1)^\intercal, \ldots, D_{1T}(\beta_1)^\intercal\}^\intercal$ and $D_2(\gamma_1) = \{D_{21}(\gamma_1)^\intercal, \ldots, D_{2T}(\gamma_1)^\intercal\}^\intercal$.

The derivative of the estimating equations $\tilde{U}(\beta_1, \gamma_1)$ and its limit as $n \to \infty$ are

$$\tilde{D}(\beta_1, \gamma_1) = \begin{pmatrix} \tilde{D}_1(\beta_1) & 0_{T\times p_1} \\ 0_{T\times p_1} & \tilde{D}_2(\gamma_1) \end{pmatrix}, D(\beta_1, \gamma_1) = \begin{pmatrix} D_1(\beta_1) & 0_{T\times p_1} \\ 0_{T\times p_1} & D_2(\gamma_1) \end{pmatrix}.$$

For the estimator $(\hat{\beta}_1^A, \hat{\gamma}_1^A)$, we have $\sqrt{n}W_1\tilde{U}(\beta_1^0, \gamma_1^0) = n^{-1/2}\sum_{i=1}^n W_1 C_i + o_p(1)$. Then $\sqrt{n}((\hat{\beta}_1^A - \beta_1^0)^\intercal, (\hat{\gamma}_1^A - \gamma_1^0)^\intercal)^\intercal$ converges in distribution to a normal random vector with covariance $A_1^{-1}B_1(A^\intercal)_1^{-1}$, where $A_1 = W_1 D(\beta_1^0, \gamma_1^0)$, $B_1 = W_1 \Sigma W_1^\intercal$ and Σ is the covariance matrix of the vector C_i.

For the estimator $(\hat{\beta}_1^B, \hat{\gamma}_1^B)$, minimizing $Q(\beta_1, \gamma_1) = \left\{\tilde{U}(\beta_1, \gamma_1)\right\}^\intercal W_2 \left\{\tilde{U}(\beta_1, \gamma_1)\right\}$ is equivalent to solving $\tilde{D}(\beta_1, \gamma_1)^\intercal W_2 \tilde{U}(\beta_1, \gamma_1) = 0$. Since

$$\sqrt{n}\tilde{D}(\beta_1^0, \gamma_1^0)^\intercal W_2 \tilde{U}(\beta_1^0, \gamma_1^0) = n^{-1/2}\sum_{i=1}^n D(\beta_1^0, \gamma_1^0)^\intercal W_2 C_i + o_p(1),$$

we conclude $\sqrt{n}((\hat{\beta}_1^B - \beta_1^0)^\intercal, (\hat{\gamma}_1^B - \gamma_1^0)^\intercal)^\intercal$ converges in distribution to a normal random vector with covariance $A_2^{-1}B_2(A_2^\intercal)^{-1}$, where $B_2 = D(\beta_1^0, \gamma_1^0)^\intercal W_2 \Sigma W_2^\intercal D(\beta_1^0, \gamma_1^0)$ and $A_2 = D(\beta_1^0, \gamma_1^0)^\intercal W_2 D(\beta_1^0, \gamma_1^0)$.

Bibliography

[1] ActiGraph, LLC, "What are counts?" https://help.theactigraph.com/entries/20723176-What-are-counts-, 12 2011, [Online; accessed 12-Feb-2015].

[2] ActiGraph, LLC, "Idle sleep mode explained," https://help.theactigraph.com/entries/21625711-Idle-Sleep-Mode-Explained, 6 2012, [Online; accessed 12-Feb-2015].

[3] ActiGraph, LLC, "Low frequency extension explained," https://help.theactigraph.com/entries/21767838-Low-Frequency-Extension-Explained, 7 2012, [Online; accessed 12-Feb-2015].

[4] M. Aittasalo, H. Vähä-Ypyä, T. Vasankari, P. Husu, A.-M. Jussila, and H. Sievänen, "Mean amplitude deviation calculated from raw acceleration data: a novel method for classifying the intensity of adolescents' physical activity irrespective of accelerometer brand," *BMC Sports Science, Medicine and Rehabilitation*, vol. 7, no. 1, p. 18, dec 2015.

[5] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. Pollak,

"The role of actigraphy in the study of sleep and circadian rhythms. American Academy of Sleep Medicine Review Paper," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.

[6] A. A. Atienza and A. C. King, "Comparing self-reported versus objectively measured physical activity behavior: A preliminary investigation of older Filipino American Women," *Research Quarterly for Exercise & Sport*, vol. 76, no. 3, pp. 358–362, 2005.

[7] J. Bai, "Accelerometer-based prediction of activity for epidemiological research," Master's thesis, Johns Hopkins University, 2011.

[8] J. Bai, C. Di, L. Xiao, K. R. Evenson, A. Z. LaCroix, C. M. Crainiceanu, and D. M. Buchner, "An Activity Index for raw accelerometry data and its comparison with other activity metrics," *PLoS ONE*, vol. 11, no. 8, p. e0160644, aug 2016.

[9] J. Bai, J. Goldsmith, B. Caffo, T. A. Glass, and C. M. Crainiceanu, "Movelets: A dictionary of movement," *Electronic Journal of Statistics*, vol. 6, pp. 559–578, 2012.

[10] J. Bai, B. He, H. Shou, V. Zipunnikov, T. A. Glass, and C. M. Crainiceanu, "Normalization and extraction of interpretable metrics from raw accelerometry data," *Biostatistics*, vol. 15, no. 1, pp. 102–116, 2014.

[11] Y. Bai, G. J. Welk, Y. H. Nam, J. A. Lee, J. M. Lee, Y. Kim, N. F. Meier, and P. M. Dixon, "Comparison of consumer and research monitors under semistructured settings," *Medicine & Science in Sports & Exercise*, vol. 48, no. 1, pp. 151–158, 2016.

[12] A. Bankoski, T. B. Harris, J. J. McClain, R. J. Brychta, P. Caserotti, K. Y. Chen, D. Berrigan, R. P. Troiano, and A. Koster, "Sedentary activity associated with metabolic syndrome independent of physical activity," *Diabetes Care*, vol. 34, no. 2, pp. 497–503, 2011.

[13] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proceedings of the 2nd International Conference on Pervasive Computing*. Springer, 2004, pp. 1–17.

[14] M. L. Blood, R. L. Sack, D. C. Percy, and J. C. Pen, "A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography," *Sleep*, vol. 20, no. 6, pp. 388–395, 1997.

[15] G. Borg and H. Linderholm, "Perceived exertion and pulse rate during graded exercise in various age groups," *Acta Medica Scandinavica*, vol. 181, pp. 192–206, 1967.

[16] J. Boyle, M. Karunanithi, T. Wark, W. Chan, and C. Colavitti, "Quantifying functional mobility progress for chronic disease management," in *Annual*

International Conference of the IEEE Engineering in Medicine and Biology - Proceedings, New York, 2006, pp. 5916–5919.

[17] S. Brage, N. Brage, P. W. Franks, U. Ekelund, and N. J. Wareham, "Reliability and validity of the combined heart rate and movement sensor Actiheart." *European Journal of Clinical Nutrition*, vol. 59, no. 4, pp. 561–570, 2005.

[18] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, "SHIMMERTM - A wireless sensor platform for noninvasive biomedical research," *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1527–1534, 2010.

[19] J. B. J. Bussmann, W. L. J. Martens, J. H. M. Tulen, F. C. Schasfoort, H. J. G. van den Berg-Emons, and H. J. Stam, "Measuring daily behavior using ambulatory accelerometry: The Activity Monitor," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 3, pp. 349–356, aug 2001.

[20] N. F. Butte, W. W. Wong, J. S. Lee, A. L. Adolph, M. R. Puyau, and I. F. Zakeri, "Prediction of energy expenditure and physical activity in preschoolers." *Medicine & Science in Sports & Exercise*, vol. 46, no. 6, pp. 1216–1226, 2014.

[21] K. L. Cain, T. L. Conway, M. A. Adams, L. E. Husak, and J. F. Sallis, "Comparison of older and newer generations of ActiGraph accelerometers with the normal filter and the low frequency extension," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 10, no. 1, p. 51, 2013.

[22] L. Choi, Z. Liu, C. E. Matthews, and M. S. Buchowski, "Validation of accelerometer wear and nonwear time classification algorithm," *Medicine & Science in Sports & Exercise*, vol. 43, no. 2, pp. 357–364, 2011.

[23] L. Choi, S. C. Ward, J. F. Schnelle, and M. S. Buchowski, "Assessment of wear/nonwear time classification algorithms for triaxial accelerometer," *Medicine & Science in Sports & Exercise*, vol. 44, no. 10, pp. 2009–2016, oct 2012.

[24] R. C. Colley, D. Garriguet, I. Janssen, C. L. Craig, J. Clarke, and M. S. Tremblay, "Physical activity of Canadian adults: accelerometer results from the 2007 to 2009 Canadian Health Measures Survey," *Health Reports*, vol. 22, no. 1, pp. 15–23, mar 2011.

[25] S. E. Crouter, J. I. Flynn, and D. R. Bassett, "Estimating physical activity in youth using a wrist accelerometer," *Medicine & Science in Sports & Exercise*, vol. 47, no. 5, pp. 944–951, 2015.

[26] M. del Rosario, S. Redmond, and N. Lovell, "Tracking the evolution of smartphone sensing for monitoring human movement," *Sensors*, vol. 15, no. 8, pp. 18 901–18 933, 2015.

[27] R. K. Dishman, R. A. Washburn, and D. A. Schoeller, "Measurement of physical activity," *Quest*, vol. 53, no. 3, pp. 295–309, aug 2001.

[28] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random

forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiological Measurement*, vol. 35, no. 11, pp. 2191–2203, 2014.

[29] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[30] D. W. Esliger, A. V. Rowlands, T. L. Hurst, M. Catt, P. Murray, and R. G. Eston, "Validation of the GENEA Accelerometer," *Medicine & Science in Sports & Exercise*, vol. 43, no. 6, pp. 1085–1093, jun 2011.

[31] K. R. Evenson, F. Wen, A. H. Herring, C. Di, M. J. LaMonte, L. F. Tinker, I.-M. Lee, E. Rillamas-Sun, A. Z. LaCroix, and D. M. Buchner, "Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91years: The Women's Health Initiative OPACH Calibration Study," *Preventive Medicine Reports*, vol. 2, pp. 750–756, 2015.

[32] A. R. Feinstein, B. R. Josephy, and C. K. Wells, "Scientific and clinical problems in indexes of functional disability," *Annals of Internal Medicine*, vol. 105, no. 3, p. 413, sep 1986.

[33] Y. Feito, H. R. Garner, and D. R. Bassett, "Evaluation of ActiGraph's low-

frequency filter in laboratory and free-living environments." *Medicine & Science in Sports & Exercise*, vol. 47, no. 1, pp. 211–217, 2015.

[34] L. Ferrucci, "The Baltimore Longitudinal Study of Aging (BLSA): a 50-year-long journey and plans for the future." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 63, no. 12, pp. 1416–9, 2008.

[35] P. S. Freedson, E. Melanson, and J. Sirard, "Calibration of the Computer Science and Applications, Inc. accelerometer," *Medicine & Science in Sports & Exercise*, vol. 30, no. 5, pp. 777–781, 1998.

[36] P. S. Freedson, D. M. Pober, and K. F. Janz, "Calibration of accelerometer output for children," *Medicine & Science in Sports & Exercise*, vol. 37, no. 11 Suppl, pp. S523—-S530, 2005.

[37] J. Goldsmith, V. Zipunnikov, and J. Schrack, "Generalized multilevel function-on-scalar regression and principal component analysis," *Biometrics*, vol. 71, no. 2, pp. 344–353, 2015.

[38] S. W. Gorny and J. R. Spiro, "Comparing different methodologies used in wrist Actigraphy," *Sleep Review*, vol. Summer, pp. 40–42, 2001.

[39] P. M. Grant, C. G. Ryan, W. W. Tigbe, and M. H. Granat, "The validation of a novel activity monitor in the measurement of posture and motion during

everyday activities," *British Journal of Sports Medicine*, vol. 40, no. 12, pp. 992–997, 2006.

[40] P. M. Grant, P. M. Dall, S. L. Mitchell, and M. H. Granat, "Activity-monitor accuracy in measuring step number and cadence in community-dwelling older adults," *Journal of Aging and Physical Activity*, vol. 16, no. 2, pp. 201–214, 2008.

[41] L. P. Hansen, "Large sample properties of Generalized Method of Moments estimators," *Econometrica*, vol. 50, no. 4, p. 1029, jul 1982.

[42] B. He, J. Bai, V. V. Zipunnikov, A. Koster, P. Caserotti, B. Lange-Maia, N. W. Glynn, T. B. Harris, and C. M. Crainiceanu, "Predicting human movement with multiple accelerometers using movelets," *Medicine & Science in Sports & Exercise*, vol. 46, no. 9, pp. 1859–1866, 2014.

[43] D. Hendelman, K. Miller, C. Baggett, E. Debold, and P. Freedson, "Validity of accelerometry for the assessment of moderate intensity physical activity in the field," *Medicine & Science in Sports & Exercise*, vol. 32, no. 9 Suppl., pp. S442–S449, 2000.

[44] M. Hildebrand, V. T. Van Hees, B. H. Hansen, and U. Ekelund, "Age-group comparability of raw accelerometer output from wrist- and hip-worn monitors," *Medicine & Science in Sports & Exercise*, vol. 46, no. 9, pp. 1816–1824, 2014.

[45] I. Janssen and A. G. LeBlanc, "Systematic review of the health benefits of physical activity and fitness in school-aged children and youth," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 7, no. 1, p. 40, 2010.

[46] G. Jean-Louis, H. von Gizycki, F. Zizi, J. Fookson, A. Spielman, J. Nunes, R. Fullilove, and H. Taub, "Determination of sleep and wakefulness with the actigraph data analysis software (ADAS)," *Sleep*, vol. 19, no. 9, pp. 739–43, 1996.

[47] F. Jelinek, *Statistical Methods for Speech Recognition.* the MIT Press, 1997.

[48] D. John and P. Freedson, "ActiGraph and actical physical activity monitors: A peek under the hood," *Medicine & Science in Sports & Exercise*, vol. 44, no. SUPPL. 1, pp. S86–S89, 2012.

[49] K. Kiani, C. J. Snijders, and E. S. Gelsema, "Computerized analysis of daily life motor activity for ambulatory monitoring." *Technology and Health Care*, vol. 5, no. 4, pp. 307–18, 1997.

[50] K. Kiani, C. J. Snijders, and E. S. Gelsema, "Recognition of daily life motor activity classes using an artificial neural network," *Archives of Physical Medicine and Rehabilitation*, vol. 79, no. 2, pp. 147–154, 1998.

[51] S. Kozey-Keadle, A. Libertine, K. Lyden, J. Staudenmayer, and P. S. Freedson,

"Validation of wearable monitors for assessing sedentary behavior," *Medicine & Science in Sports & Exercise*, vol. 43, no. 8, pp. 1561–1567, 2011.

[52] A. Krause, D. P. Siewiorek, A. Smailagic, and J. Farringdon, "Unsupervised, dynamic identification of physiological and activity context in wearable computing," in *Wearable Computers, The Seventh International Symposium on.* IEEE Computer Society, 2003, pp. 88–97.

[53] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients," *Sleep Medicine*, vol. 2, no. 5, pp. 389–396, 2001.

[54] I.-M. Lee and R. S. Paffenbarger, "Associations of light, moderate, and vigorous intensity physical activity with longevity. The Harvard Alumni Health Study." *American Journal of Epidemiology*, vol. 151, no. 3, pp. 293–299, 2000.

[55] H. Li, S. Kozey Keadle, J. Staudenmayer, H. Assaad, J. Z. Huang, and R. J. Carroll, "Methods to assess an exercise intervention trial based on 3-level functional data," *Biostatistics*, vol. 16, no. 4, pp. 754–771, oct 2015.

[56] H. Li, J. Staudenmayer, and R. J. Carroll, "Hierarchical functional data with mixed continuous and binary measurements," *Biometrics*, vol. 70, no. 4, pp. 802–811, 2014.

[57] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 2. IEEE Press, 2001.

[58] I. McDowell and C. Newell, *Measuring Health: A Guide to Rating Scales and Questionnaires*. New York: Oxford University Press, 1987.

[59] K. Mishima, Y. Hishikawa, and M. Okawa, "Randomized, dim light controlled, crossover test of morning bright light therapy for rest-activity rhythm disorders in patients with vascular dementia and dementia of Alzheimer's type," *Chronobiology International*, vol. 15, no. 6, pp. 647–654, 1998.

[60] H. K. Neilson, P. J. Robson, C. M. Friedenreich, and I. Csizmadi, "Estimating activity energy expenditure: How valid are physical activity questionnaires?" *American Journal of Clinical Nutrition*, vol. 87, no. 2, pp. 279–291, 2008.

[61] M. E. Nelson, W. J. Rejeski, S. N. Blair, P. W. Duncan, J. O. Judge, A. C. King, C. A. Macera, and C. Castaneda-Sceppa, "Physical activity and public health in older adults: Recommendation from the American College of Sports Medicine and the American Heart Association," *Circulation*, vol. 116, no. 9, pp. 1094–1105, aug 2007.

[62] A. Nguyen, D. Moore, and I. McCowan, "Unsupervised clustering of free-living human activities using ambulatory accelerometry," *Annual International Con-*

ference of the IEEE Engineering in Medicine and Biology - Proceedings, pp. 4895–4898, 2007.

[63] K. J. O'Donovan, B. R. Greene, D. McGrath, R. O'Neill, A. Burns, and B. Caulfield, "SHIMMER: A new tool for temporal gait analysis," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.* IEEE, 2009, pp. 3826–3829.

[64] J. Pärkkä, M. Ermes, P. Korpipää, J. Mäntyjärvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[65] R. R. Pate, M. Pratt, S. N. Blair, W. L. Haskell, C. A. Macera, C. Bouchard, D. Buchner, W. Ettinger, G. W. Heath, A. C. King, A. Kriska, A. S. Leon, B. H. Marcus, J. Morris, R. S. Paffenbarger, K. Patrick, M. L. Pollock, J. M. Rippe, J. Sallis, and J. H. Wilmore, "Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine," *Journal of the American Medical Association*, vol. 273, no. 5, pp. 402–407, 1995.

[66] Z. Pedisic and A. Bauman, "Accelerometer-based measures in physical activity surveillance: current practices and issues," *British Journal of Sports Medicine*, vol. 49, no. 4, pp. 219–223, 2015.

[67] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, vol. 81, no. 9. IEEE Press, 1993, pp. 1215–1247.

[68] M. Pinsky and S. Karlin, *An Introduction to Stochastic Modeling*, 4th ed. Academic Press, 2010.

[69] D. M. Pober, J. Staudenmayer, C. Raphael, and P. S. Freedson, "Development of novel techniques to classify physical activity mode using accelerometers," *Medicine & Science in Sports & Exercise*, vol. 38, no. 9, pp. 1626–1634, 2006.

[70] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors–a review of classification techniques," *Physiological Measurement*, vol. 30, no. 4, pp. R1–R33, 2009.

[71] S. A. Prince, K. B. Adamo, M. Hamel, J. Hardt, S. Connor Gorber, and M. Tremblay, "A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 5, no. 1, p. 56, 2008.

[72] M. R. Puyau, A. L. Adolph, F. A. Vohra, and N. F. Butte, "Validation and calibration of physical activity monitors in children," *Obesity Research*, vol. 10, no. 3, pp. 150–157, 2002.

[73] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from

accelerometer data," in *Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence*, vol. 5, 2005, pp. 1541–1546.

[74] J. F. Sallis, B. E. Saelens, L. D. Frank, T. L. Conway, D. J. Slymen, K. L. Cain, J. E. Chapman, and J. Kerr, "Neighborhood built environment and income: Examining multiple health outcomes," *Social Science & Medicine*, vol. 68, no. 7, pp. 1285–1293, apr 2009.

[75] J. E. Sasaki, D. John, and P. S. Freedson, "Validation and comparison of ActiGraph activity monitors," *Journal of Science and Medicine in Sport*, vol. 14, no. 5, pp. 411–416, 2011.

[76] J. A. Schrack, V. Zipunnikov, J. Goldsmith, J. Bai, E. M. Simonsick, C. Crainiceanu, and L. Ferrucci, "Assessing the "Physical Cliff": Detailed quantification of age-related differences in daily patterns of physical activity," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 69, no. 8, pp. 973–979, 2014.

[77] J. A. Schrack, V. Zipunnikov, J. Goldsmith, K. Bandeen-Roche, C. M. Crainiceanu, and L. Ferrucci, "Estimating energy expenditure from heart rate in older adults: A case for calibration," *PLoS ONE*, vol. 9, no. 4, p. e93520, 2014.

[78] *Shimmer 9DoF Calibration Application User Manual*, http://www.shimmer-research.com/wp-content/uploads/2012/02/

Shimmer-9DoF-Calibration-User-Manual.pdf, Shimmer Research, 2012, v 1.0b.

[79] H. Shou, V. Zipunnikov, C. M. Crainiceanu, and S. Greven, "Structured functional principal component analysis," *Biometrics*, vol. 71, no. 1, pp. 247–257, mar 2015.

[80] J. Staudenmayer, S. He, A. Hickey, J. Sasaki, and P. Freedson, "Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements," *Journal of Applied Physiology*, vol. 119, no. 4, pp. 396–403, aug 2015.

[81] J. Staudenmayer, D. Pober, S. Crouter, D. Bassett, and P. Freedson, "An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer," *Journal of Applied Physiology*, vol. 107, no. 4, pp. 1300–1307, 2009.

[82] J. Symanzik and W. Shannon, "Exploratory graphics for functional Actigraphy data," in *Joint Statistical Meeting 2008*, 2008, pp. 3707–3714.

[83] D. J. Terbizan, B. a. Dolezal, and C. Albano, "Validity of seven commercially available heart rate monitors," *Measurement in Physical Education and Exercise Science*, vol. 6, no. 4, pp. 243–247, dec 2002.

[84] M. S. Treuth, K. Schmitz, D. J. Catellier, R. G. McMurray, D. M. Murray,

M. J. Almeida, S. Going, J. E. Norman, and R. Pate, "Defining accelerometer thresholds for activity intensities in adolescent girls." *Medicine & Science in Sports & Exercise*, vol. 36, no. 7, pp. 1259–66, 2004.

[85] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Mâsse, T. Tilert, and M. McDowell, "Physical activity in the United States measured by accelerometer," *Medicine & Science in Sports & Exercise*, vol. 40, no. 1, pp. 181–188, 2008.

[86] R. P. Troiano, J. J. McClain, R. J. Brychta, and K. Y. Chen, "Evolution of accelerometer methods for physical activity research," *British Journal of Sports Medicine*, vol. 48, no. 13, pp. 1019–1023, jul 2014.

[87] S. G. Trost, R. R. Pate, J. F. Sallis, P. S. Freedson, W. C. Taylor, M. Dowda, and J. Sirard, "Age and gender differences in objectively measured physical activity in youth," *Medicine & Science in Sports & Exercise*, vol. 34, no. 2, pp. 350–355, 2002.

[88] J. K. Urbanek, J. Harezlak, N. W. Glynn, T. Harris, C. Crainiceanu, and V. Zipunnikov, "Stride variability measures derived from wrist- and hip-worn accelerometers," *Gait & Posture*, vol. 52, pp. 217–223, feb 2017.

[89] H. Vähä-Ypyä, T. Vasankari, P. Husu, A. Mänttäri, T. Vuorimaa, J. Suni, and H. Sievänen, "Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based Mean Amplitude Deviation (MAD)," *PLOS ONE*, vol. 10, no. 8, p. e0134813, aug 2015.

[90] V. T. van Hees, L. Gorzelniak, E. C. Dean León, M. Eder, M. Pias, S. Taherian, U. Ekelund, F. Renström, P. W. Franks, A. Horsch, and S. Brage, "Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity," *PLoS ONE*, vol. 8, no. 4, p. e61691, apr 2013.

[91] G. J. Welk, "Principles of design and analyses for the calibration of accelerometry-based activity monitors," *Medicine & Science in Sports & Exercise*, vol. 37, no. 11 SUPPL., pp. 501–511, 2005.

[92] G. J. Welk, S. N. Blair, K. Wood, S. Jones, and R. W. Thompson, "A comparative evaluation of three accelerometry-based physical activity monitors," *Medicine & Science in Sports & Exercise*, vol. 32, no. 9; SUPP/1, pp. S489–S497, 2000.

[93] L. Xiao, L. Huang, J. A. Schrack, L. Ferrucci, V. Zipunnikov, and C. M. Crainiceanu, "Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach," *Biostatistics*, vol. 16, no. 2, pp. 352–367, apr 2015.

[94] C. C. Yang and Y. L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010.

[95] M. Yuwono, S. W. Su, Y. Guo, B. D. Moulton, and H. T. Nguyen, "Unsuper-

vised nonparametric method for gait analysis using a waist-worn inertial sensor," *Applied Soft Computing*, vol. 14, pp. 72–80, jan 2014.

[96] K. Zhang, F. X. Pi-Sunyer, and C. N. Boozer, "Improving energy expenditure estimation for physical activity," *Medicine & Science in Sports & Exercise*, vol. 36, no. 5, pp. 883–889, 2004.

[97] K. Zhang, P. Werner, M. Sun, F. X. Pi-Sunyer, and C. N. Boozer, "Measurement of human daily physical activity," *Obesity Research*, vol. 11, no. 1, pp. 33–40, 2003.

Vita

Jiawei Bai, born on April 17, 1987 in Chengdu, Sichuan, China, received his Bachelor of Science in Statistics from Sichuan University in 2009. He moved to United Stated thereafter and enrolled in the Master of Science program in Biostatistics at the Johns Hopkins University. He received his Master of Science degree in 2011 and started his Doctoral of Philosophy study at the same department in 2012. His research in statistical methods and applications of wearable devices has won him awards including the Jane and Steve Dykacz Award (2013), the Louis I. and Thomas D. Dublin Award (2014), ENAR Regional Advisory Board Poster Award (2016), ENAR Distinguished Student Paper Award (2017) and the Margaret Merrell Award (2017).

CURRICULUM VITAE

Jiawei BAI

EDUCATION

2012 – 2017	**Doctor of Philosophy, Biostatistics** Johns Hopkins University, Baltimore, Maryland, USA Thesis: Statistical methods for wearable devices with applications to epidemiological studies Advisor: Dr. Ciprian Crainiceanu, Co-advisor: Dr. Vadim Zipunnikov
2009 – 2011	**Master of Science, Biostatistics** Johns Hopkins University, Baltimore, Maryland, USA Thesis: Accelerometer-based prediction of activity for epidemiological research Advisor: Dr. Ciprian Crainiceanu
2005 – 2009	**Bachelor of Science, Statistics (Honors Program)** Sichuan University, Chengdu, China Thesis: A study of integrated risk measurement based on Archimedean copula Advisor: Dr. Zhuyu Li

PROFESSIONAL EXPERIENCE

2009 Fall – 2017 Spring	**Research Assistant** Department of Biostatistics, JHSPH
2016 Summer & 2015 Fall	**Visiting Graduate Student** Fred Hutchinson Cancer Research Center
2015 Summer	**Summer Research Intern** AT&T Labs - Middletown
2014 Summer	**Summer Research Intern** AT&T Labs - Middletown

HONORS & AWARDS

2017	**Margaret Merrell Award** For outstanding research by a Biostatistics doctoral student JHSPH, 2017

2017	**ENAR Distinguished Student Paper Award** ENAR, 2017
2016	**ENAR Regional Advisory Board Poster Award** ENAR Regional Advisory Board, 2016
2014	**The Louis I. and Thomas D. Dublin Award** Best student research focusing on the effective use of statistical reasoning and methods in epidemiology JHSPH, 2014
2013	**The Jane and Steve Dykacz Award** Best paper by a Biostatistics student in the area of medical statistics JHSPH, 2013

PUBLICATIONS

1. **J. Bai**, A. Ivanescu and C. M. Crainiceanu. Discussion of the paper "A general framework for Functional Regression Modelling", *Statistical Modelling*, 2017; Volume 17 (1-2): 36-44.
2. **J. Bai**, C. Di, L. Xiao, K. R. Evenson, A. Z. LaCroix, C. M. Crainiceanu and D. M. Buchner. An Activity Index for raw accelerometry data and its comparison with other activity metrics. *PLoS ONE* 11(8): e0160644.
3. B. He, **J. Bai**, V. Zipunnikov, A. Koster, P. Caserotti, B. Lange-Maia, N. W. Glynn, T. B. Harris and C. M. Crainiceanu. Predicting human movement with multiple accelerometers using Movelets, *Medicine & Science in Sports & Exercise*, 2014 Sep; Volume 46 (9): 1859-1866.
4. J. A. Schrack, V. Zipunnikov, J. Goldsmith, **J. Bai**, E. M. Simonsick, C. M. Crainiceanu, L. Ferrucci. Assessing the "Physical Cliff:" Detailed quantification of aging and physical activity, *Journal of Gerontology: Medical Sciences*, 2014; Volume 69 (8): 973-979.
5. **J. Bai**, B. He, H. Shou, V. Zipunnikov, T. A. Glass, and C. M. Crainiceanu. Normalization and extraction of interpretable metrics from raw accelerometry data, *Biostatistics*, 2014 Jan; Volume 15 (1): 102-116.
6. S. Yin, M. Jensen, **J. Bai**, C. DebRoy, R. Barrangou, and E. Dudley. Evolutionary divergence of Shiga Toxin-producing Escherichia coli is reflected in CRISPR spacer composition, *Applied and Environmental Microbiology*, 2013 Sep; Volume 79 (18): 5710-5720.
7. **J. Bai**, J. Goldsmith, B. Caffo, T. A. Glass, and C. M. Crainiceanu. Movelets: A dictionary of movement, *Electronic Journal of Statistics*, 2012; Volume 6, 559-578.

ORAL & POSTER PRESENTATIONS

1. A two-stage model for wearable device data. Oral presentation (invited). ICSA Applied Statistics Symposium 2016, Atlanta, GA, USA
2. An activity index for raw accelerometry data and its comparison with activity counts. Oral

poster (contributed). ENAR 2016, Austin, TX, USA

3. An improved version of Activity Intensity and its comparison with Activity Count, with the application to Women Health Initiative. Speed session. Joint Statistical Meeting 2015, Seattle, WA, USA
4. Characterizing types of physical activity: an unsupervised way. Oral presentation (contributed). ENAR 2015, Miami, FL, USA
5. Movelet method & activity prediction. PennSIVE Research Group Meeting. Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
6. "Making use of the type of activity" in workshop "Daily activity in old age: New ways to capture function in the community". Pre-conference workshop lecture (invited). Gerontological Society of America 2014 Annual Scientific Meeting, Washington, DC, USA
7. Using movelets and its extensions to characterize types of physical activity. Oral presentation (topic contributed). Joint Statistical Meeting 2014, Boston, MA, USA
8. Quantifying circadian trajectory of fatigability using the proportional intensity model. Oral presentation (contributed). ENAR 2014, Baltimore, MD, USA
9. Two valid and interpretable metrics to summarize raw accelerometry data. Oral presentation (contributed). Joint Statistical Meeting 2013, Montreal, QC, Canada
10. Two valid and interpretable metrics to summarize raw accelerometry data. Poster presentation (contributed). 28th International Workshop on Statistical Modelling, Palermo, Italy
11. Normalization and extraction of interpretable metrics from raw accelerometry data. Oral presentation (student competition finalist). International Conference on Ambulatory Monitoring of Physical Activity and Movement 2013, Amherst, MA, USA
12. Accelerometry Metrics for Epidemiology. Oral presentation (contributed). ENAR 2013, Orlando, FL, USA
13. Movelets: A dictionary of movement. Oral presentation (contributed). JSM 2012, San Diego, CA, USA
14. Accelerometer-based prediction of activity for epidemiological research. Oral presentation (contributed). JSM 2011, Miami, FL, USA

SOFTWARE

ActivityIndex	An R package which provides functions to read and process raw accelerometry data. Link: https://github.com/javybai/ActivityIndex

TEACHING EXPERIENCE

2017 Spring	**Lead Teaching Assistant**: [140.653-654] Methods in Biostatistics III & IV
2016 Spring	**Lead Teaching Assistant**: [140.653-654] Methods in Biostatistics III & IV

2015 Spring	**Teaching Assistant**: [140.653-654] Methods in Biostatistics III & IV
2014 Spring	**Teaching Assistant**: [140.653-654] Methods in Biostatistics III & IV
2013 Spring	**Teaching Assistant**: [140.623-624] Statistical Methods in Public Health III & IV
2012 Fall	**Teaching Assistant**: [140.621-622] Statistical Methods in Public Health I & II
2011 Spring	**Teaching Assistant**: [140.623-624] Statistical Methods in Public Health III & IV
2010 Fall	**Teaching Assistant**: [AS.280.345] Biostatistics in Public Health

PROFESSIONAL ACTIVITIES

2016 Spring	**Co-organizer, Epidemiology and Biostatistics of Aging Progress Showcase** JHSPH
2013 – 2014	**Co-organizer, Student Computing Club** Department of Biostatistics, JHSPH
2013 – 2014	**PhD Student Representative** Department of Biostatistics, JHSPH
2013 – 2014	**Student Assembly Departmental Representative** JHSPH
2010 – 2012	**Committee Member for Technology & Logistics** Chinese Public Health Forum at JHSPH